

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE



REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified		1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.	
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE		4. PERFORMING ORGANIZATION REPORT NUMBER(S) MIT/LCS/TR 509	
5. MONITORING ORGANIZATION REPORT NUMBER(S) N00014-82-K-0727/ N00014-89-J-1332		6a. NAME OF PERFORMING ORGANIZATION MIT Lab for Computer Science	
6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION Office of Naval Research/Dept. of Navy	
6c. ADDRESS (City, State, and ZIP Code) 545 Technology Square Cambridge, MA 02139		7b. ADDRESS (City, State, and ZIP Code) Information Systems Program Arlington, VA 22217	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION DARPA/DOD		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER	
8c. ADDRESS (City, State, and ZIP Code) 1400 Wilson Blvd. Arlington, VA 22217		10. SOURCE OF FUNDING NUMBERS	
PROGRAM ELEMENT NO.		PROJECT NO.	TASK NO.
WORK UNIT ACCESSION NO.			
11. TITLE (Include Security Classification) An Information-Theoretical Approach to Studying Phoneme Collocational Constraints			
12. PERSONAL AUTHOR(S) Robert Howard Kassel			
13a. TYPE OF REPORT Technical	13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) July 1991	15. PAGE COUNT 70
16. SUPPLEMENTARY NOTATION			
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	
19. ABSTRACT (Continue on reverse if necessary and identify by block number)			
<p>This thesis describes a lexical study of phoneme collocational constraints using a metric motivated by information theory. Phonologists have long been describing the permissible combination of phonemes in the form of phonotactic rules. They have shown that these rules often can be expressed in terms of phoneme equivalence classes. Thus, for example, the homorganic rule for American English states the a syllable-final nasal-stop pair must agree on their place of articulation. Over the past decade, there have also been many lexical studies examining the constraining power of phoneme equivalence classes, demonstrating their utility for lexical access. While there are good reasons to express these constraints using classes well motivated by theory, the phoneme space clearly can be partitioned in many other ways. It is conceivable that, by allowing phonemes to form various sets of equivalence classes and quantifying the constraining power for each set, we may discover phoneme classes that will provide the strongest constraints for lexical access. Our line of investigation is inspired by recent work on word equivalence classes by Jelinek and word collocational constraints by Church.</p>			
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS		21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Carol Nicolora		22b. TELEPHONE (Include Area Code) (617) 253-5894	22c. OFFICE SYMBOL

19.

Specifically, we investigated phoneme collocational constraints using a normalized measure of mutual information. A pair-wise, hierarchical clustering technique is used to combine phonemes into classes using this metric. The result of this clustering procedure can be displayed as a dendrogram, from which an arbitrary number of equivalence classes can be selected.

We have conducted a number of experiments investigating the collocation constraints of phoneme pairs and triplets. We found that in many cases phonemes are organized into classes that share certain phonological features. In fact, phonemes that have similar *acoustic* properties often exhibit similar collocational constraints. We also compared the constraining power of our phoneme classes with those chosen with a phonological criterion, and found ours to be more than competitive. Based on our results, we conclude that our information theoretic metric is particularly well suited to a description of lexical constraining power. We discuss the implications of the results to automatic speech recognition.



Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

*In memory of
my Dad.*

An Information-Theoretical Approach
to Studying Phoneme Collocational Constraints

by

Robert Howard Kassel
S.B., Massachusetts Institute of Technology
(1986)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements
for the degree of

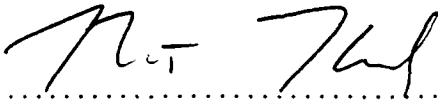
Master of Science

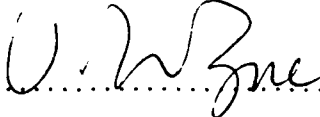
at the

Massachusetts Institute of Technology
June, 1990

©Robert Kassel and the Massachusetts Institute of Technology, 1990.
All rights reserved.

The author hereby grants to MIT permission to reproduce
and to distribute copies of this thesis document
in whole or in part.

Signature of Author 
Department of Electrical Engineering and Computer Science
May, 1990

Certified by 
Victor W. Zue
Principal Research Scientist,
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by
Arthur C. Smith
Chair, Department Committee on Graduate Students

91-06838


An Information-Theoretical Approach
to Studying Phoneme Collocational Constraints

by

Robert H. Kassel

Submitted to the Department of Electrical Engineering and Computer Science
in May, 1990 in partial fulfillment of the requirements for the degree of
Master of Science.

Abstract

This thesis describes a lexical study of phoneme collocational constraints using a metric motivated by information theory. Phonologists have long been describing the permissible combination of phonemes in the form of phonotactic rules. They have shown that these rules often can be expressed in terms of phoneme equivalence classes. Thus, for example, the homorganic rule for American English states that a syllable-final nasal-stop pair must agree on their place of articulation. Over the past decade, there have also been many lexical studies examining the constraining power of phoneme equivalence classes, demonstrating their utility for lexical access. While there are good reasons to express these constraints using classes well motivated by theory, the phoneme space clearly can be partitioned in many other ways. It is conceivable that, by allowing phonemes to form various sets of equivalence classes and quantifying the constraining power for each set, we may discover phoneme classes that will provide the strongest constraints for lexical access. Our line of investigation is inspired by recent work on word equivalence classes by Jelinek and word collocational constraints by Church.

Specifically, we investigated phoneme collocational constraints using a normalized measure of mutual information. A pair-wise, hierarchical clustering technique is used to combine phonemes into classes using this metric. The result of this clustering procedure can be displayed as a dendrogram, from which an arbitrary number of equivalence classes can be selected.

We have conducted a number of experiments investigating the collocation constraints of phoneme pairs and triplets. We found that in many cases phonemes are organized into classes that share certain phonological features. In fact, phonemes that have similar *acoustic* properties often exhibit similar collocational constraints. We also compared the constraining power of our phoneme classes with those chosen with a phonological criterion, and found ours to be more than competitive. Based on our results, we conclude that our information theoretic metric is particularly well suited to a description of lexical constraining power. We discuss the implications of the results to automatic speech recognition.

Thesis Supervisor: Victor W. Zue

Title: Principal Research Scientist

Acknowledgements

To Victor Zue, my thesis advisor, for helping me to focus on this project and for providing an outstanding research environment;

To Nancy Daly, Jim Glass, and Stephanie Seneff, for reading drafts of this thesis and providing many helpful comments;

To my family, especially Jason, for giving me so much support and understanding;

To Dave Whitney, for friendship and sanity and knee-deep snow and jumping out of a "perfectly good plane;"

To Lori Lamel, for listening and encouraging, and for a bottle of motivation I hope to open soon;

To Vicky Palay, for endless assistance and cheer;

To the members of the Spoken Language Systems Group, for your contributions, too numerous to list;

To the fine gentlemen of F-Entry, for giving me a home instead of a place to live;

And to Dan Huttenlocher, for getting me started way back when;

I owe you all my deepest thanks.

This research was supported by DARPA-ISTO.

Contents

1	Introduction	9
1.1	Units and their Organization	9
1.2	Linguistic Description of Phonemic Constraints	10
1.3	Previous Computational Studies	11
1.4	Collocational Constraints in Language	12
1.5	Thesis Overview	13
2	Approach	15
2.1	General Considerations	15
2.1.1	Use a Minimum of Preconceptions	16
2.1.2	Maximize Data Utilization	16
2.1.3	Apply Information-Theoretic Measures	17
2.1.4	Relevance to Lexical Access	17
2.2	Metric Overview	18
2.2.1	Metric Development	18
2.2.2	Normalization	19
2.3	Pronunciation Constructs	20
2.3.1	Validity of Using Phonemes	20
2.3.2	Syllables are Too Controversial	21
2.3.3	Word Boundary Independence	21
2.4	Clustering Technique Overview	22
2.4.1	Many Possible Classes	23
2.4.2	Algorithm	23
2.4.3	Reducing Computational Complexity	23
2.5	Lexicon Preparation	24
2.5.1	Modifying Pronunciations	24
2.5.2	Word Frequency Weighting	25
2.6	An Example	25
2.7	Summary	29
3	Experiments	31
3.1	Diphones	31
3.2	Triphones	33
3.3	Cluster Evaluation	35

3.3.1	Lexical Experiments	35
3.3.2	Baseline Establishment	36
3.3.3	Distinctive Features	36
3.3.4	Acoustic Clustering	41
3.3.5	Results	42
3.4	Discussion	44
3.4.1	Capturing Collocational Constraints	44
3.4.2	Relationship to Pattern Recognition	45
3.4.3	Effects of Altering Pronunciations	47
3.4.4	Lexicon Idiosyncrasies	47
3.5	Summary	51
4	Conclusions	53
4.1	Summary of Results	53
4.2	Suggested Extensions	54
4.2.1	Word Sequence Modelling	54
4.2.2	Significance Pruning of Seed Sequences	54
4.2.3	Alternate Clustering Techniques	55
4.2.4	Recognizer Tuned Clusters	55
4.2.5	Acoustic Detectability	55
4.2.6	Measuring Tree Stability	55
4.3	Summary	56
A	Related Clustering Experiments	57
A.1	Directional Diphone Measure	57
A.2	Forward Prediction Triphones	58
A.3	Phoneme Class Stability	59
A.4	Word Frequency Weighting	59
A.5	Longer Range Effects	60
B	Related Lexical Experiments	63

Chapter 1

Introduction

This thesis explores phoneme collocational constraints, their use in discovering phonological equivalence classes, and their application for lexical access. Specifically, we employ an information theoretic metric over the collocational constraints to form a phoneme hierarchy. In this chapter we review how previous studies have investigated phonemic constraints. We will then discuss both the empirical and experimental bases which motivate our study.

1.1 Units and their Organization

Spoken language is composed of units of varying scales. Phonemes constitute the finite set of contrastive sounds in a language. The syllable, though controversial, seems involved in the mental representation of language. Morphemes are the smallest unit of meaning. Words, phrases, sentences, and discourse convey more detailed expressions. Units at a particular scale are combined to form the units of larger scales.

These units cannot be combined naphazardly. Language provides constraints on how the units can be assembled to form valid structures. The constraints can be applied to improve the performance of machine-based speech recognition by reducing the difficulty of the task. To do so it is important to understand how the constraints can be represented and how much constraining power they offer.

Constraints at different levels have been studied with varying degrees of thoroughness. Word level constraints, which govern ordering of words into larger units, have

been particularly well-explored. Syntactic constraints have been codified in the form of grammars. Semantic constraints, for example restrictions between verbs and their objects, are beginning to be understood. Constraining power can be estimated by established metrics like perplexity.

Our understanding of constraints at other levels is comparatively primitive and, to a large extent, anecdotal. In particular, the organization of phonemes into words is still poorly understood although its importance is unquestioned. How should these constraints be expressed? Can we quantify the applicability of a particular constraint? Is there a measure of correctness or excellence we can apply to evaluate discordant constraints?

1.2 Linguistic Description of Phonemic Constraints

Linguists have developed phonotactic rules to describe phoneme collocational constraints, but they usually are achieved through enumeration and introspection. These rules are based primarily on phoneme environments, but factors from other levels can be incorporated into the constraints as well. For example, the homorganic nasal-stop rule states that a nasal followed by a stop within a syllable must be one of the following pairs:

/mp/	/nt/	/ŋk/
/mb/	/nd/	/ŋg/

Thus we have words like /læmp/ and /lænd/ but not */lænp/. This restriction is not enforced across syllable boundaries, as in /mklud/. Knowledge of this constraint can be used to aid lexical access in a speech recognition system by helping to anchor word or syllable boundaries at non-homorganic pairs.

While these rules can be specified by enumeration of allowable sequences, they can be more compactly described using phonological properties, often properties suggested by linguists for other purposes. In the example above, we can say that the nasal and stop must have the same place of articulation. The fact that these constraints can be described in terms of properties suggests that phonemes may be organized into more than a flat structure.

One possible structuring mechanism is based on distinctive features [1]. These are a set of phonetic properties which describe a sound's manner of production and place of articulation. Sagey [2] proposes a structure in which the features are embedded in a hierarchy universal across languages. Because features can be used to partition the phoneme space into equivalence classes, it is implied that phonemes too can be arranged into a universal hierarchy.

Stevens [3] discusses how one might use distinctive features for lexical access. He argues that the acoustic correlates of distinctive features are more robust than those of a particular allophone. Features provide a compact manner of representing allophonic variation and can specify both abrupt and gradual articulator movements. Furthermore, the redundancy inherent in features suggests an underspecified representation at the lexical level. Regardless of the mechanical benefits afforded by a feature-based lexical access strategy, we should ask how much lexical constraining power they provide.

1.3 Previous Computational Studies

Because spoken language has always had to contend with communication through a noisy channel, we expect that it has evolved to include mechanisms to enhance robustness. It is these mechanisms we wish to discover and exploit.

Researchers have tried to quantify the constraining power of phonotactics. In doing so, they hope to find a set of classes which avoids the need to make fine phonetic distinction yet captures much of the constraining power inherent in the lexicon. Such broad classes are presumably more robust and easier to detect than the phonemes which they comprise.

Shipman and Zue [4] performed studies showing that even a broadly characterized phoneme string provides substantial constraints for lexical access of isolated words. In some cases, the constraint is sufficient to identify the word without finer analysis. They categorized each phoneme into one of six manner classes and used the classes to map a lexicon into cohorts containing words with the same broad class patterns. To measure the efficiency of the broad classes they computed various statistics on the cohorts' sizes. Huttenlocher [5] refined the study by incorporating a better metric and

exploring the effects of lexical stress. He also showed how acoustic detectors could be constructed for the broad classes, an idea more fully developed by Fissore, et al. [6], as part of a complete speech recognition system. However none of these studies explored the effects of varying the broad classes and there was no experimental basis for the particular classes they chose.

The lexical metrics used in these studies are questioned by Carter [7]. He argues that a good lexical access metric should use a logarithmic scale, as it better characterizes the amount of additional work needed to identify a word. Accordingly he suggests using word entropy over a lexicon's broad-class cohorts to measure the constraining power provided.

Vernooij, et al. [8] further refine this work by noting that after broad classification and lexical access we still need to perform finer phonetic classification to identify a word in a cohort. Again we would like to make as broad a categorization as possible to avoid classification errors. They use a constrained clustering technique to determine the best intermediate classes between their five broad classes and the phonemes. Their metric is based on the number of words uniquely identified by a proposed set of classes.

1.4 Collocational Constraints in Language

Previous computational studies relied on the introspection of researchers for the broad classification used. Clearly, there are many ways phonemes can be partitioned into equivalence classes. We would like to determine if some reasonable set of broad classes, reasonable in terms of existing linguistic theories and what might be detected acoustically, can be derived in a data-driven manner. If so, we will have a powerful confirmation that our intuition is right. If not, we can at least use the results to gauge the relative constraining power of some other broad class set.

Work by Shannon [9] has shown that there are strong constraints on letter sequences which can be applied to efficiently encoding texts. He demonstrated this by applying an information-theoretic metric to letter strings. As the length of the string increases, the uncertainty of the following letter decreases. Unfortunately, longer letter sequences also capture more idiosyncrasies of the text studied and so are less

applicable to other texts.

Studies by Jelinek [10] have shown that an information theoretic metric can be used to determine word categories from corpora. By combining words which occur in similar contexts, classes embodying both syntactic and semantic information are formed. These language structures are captured automatically without resorting to the incorporation of syntax or semantics in the metric. Church, et al., [11] perform similar computations but relax the word ordering requirement. This compensates for noise in the form of inserted words.

Our goal is to apply similar techniques to pronunciations in order to find phoneme classes. In doing so, we will provide a more complete analysis of phoneme classes than is present in other studies. We are encouraged by the success of word-level studies as they capture linguistic information without explicitly requiring it. We expect to utilize the analogous structures at the phoneme level.

1.5 Thesis Overview

In this thesis, we will demonstrate a technique for capturing phoneme collocational constraints by classifying phonemes into a hierarchy. The approach we take is data-driven and relies on large lexicons to represent the language. By using an information-theoretic metric we will provide results which are meaningful and match the lexical access task's complexity. We evaluate these classes against classes suggested by other studies using measures motivated by the lexical access task. We propose a new evaluation metric which may be better suited to recognition systems, particularly continuous speech systems, than previous measures.

The remainder of this thesis is as follows: In chapter 2, we outline the issues important to our study and give our philosophy for addressing them. Chapter 3 shows how we automatically derive phoneme equivalence classes using large lexica, and provides comparison using historical measures. Finally, in chapter 4 we discuss possible extensions of this work.

Chapter 2

Approach

In this chapter we will outline our philosophy and methods of examining phoneme collocational constraints. We begin by looking at the general requirements and how we can avoid the weaknesses in previous work. Next, we provide the information theoretic measures used by this study and explain the issues involved in using them on phoneme strings. We then discuss the clustering technique and lexicon preparation. Finally, we give an example of how we can use collocational constraints to form phoneme classes.

2.1 General Considerations

We believe that a major flaw with previous phoneme and lexical studies were the preconceived notions as to how phonemes should be classified. There are proposed linguistic and acoustic classification schemes, but they may not entirely address the needs of lexical access specifically.

Vernooij, et al., demonstrated using a self-organizing technique to develop a hierarchy spanning from broad-classes to phonemes that is oriented to speech recognition systems. However, they did not carry the experiments to their logical conclusion. Their technique can be used to organize all phonemes into a single hierarchy, a hierarchy free of the somewhat arbitrary broad classes. Then the entire hierarchy, including a new set of broad classes, will be “optimized” for lexical access. We would expect a system using such a hierarchy to deliver better lexical access performance than one in which part of the hierarchy is selected according to other criteria.

2.1.1 Use a Minimum of Preconceptions

For lexical experiments we must choose a particular representation for pronunciations; in particular we must specify a phoneme inventory. A good metric should treat the resulting data symbolically and make no further assumptions about what sounds the symbols represent or how to organize them. Although we know linguistic units larger than the phonemes play a role in phonotactic constraints, we do not fully understand this mechanism. For both of these reasons, we place no additional constraints on the organization of phonemes. We make no claim as to the acoustic similarity of the phonemes in a cluster, nor do we require the cluster to conform to an external set of linguistic criteria.

2.1.2 Maximize Data Utilization

Phonotactics provide powerful constraints on sound patterns. For example, we know that there are severe restrictions on the permissible consonant clusters in English. A good metric should be able to exploit these constraints to avoid relying on other sources. A metric based on word comparisons does not adequately do this.

Consider two words, one of the form **CVCC** and the other **CCVC** where **C** represents a consonant and **V** represents a vowel. A metric which uses strict string comparison compares the n^{th} characters of the two strings. In this example we compare vowels against consonants while a better method might compare the clusters and vowels separately. To remedy this we could try performing an alignment of sorts between the pronunciations. It is not clear how to carry out such an alignment. Furthermore, the alignment process would inherently impart a bias to the results.

There is an additional problem with this measure: it partitions the lexicon based on the number of phonemes in each pronunciation. Thus it only compares a word against words with a like number of phonemes. This can result in sparse data problems and may make us miss important comparisons. It also implies that our lexical access strategy needs to compare only words of equal pronunciation length. This is only true if our recognition system always hypothesizes the correct number of phonemes.

Instead we would prefer a metric which compares a word against all others. A metric based on phoneme collocational data can do this because it represents all

pronunciations in an abstract form. By using such a metric we in turn hope to develop a phoneme hierarchy which reflects the phoneme collocational constraints.

2.1.3 Apply Information-Theoretic Measures

Other studies have demonstrated the power of applying information measures in other language related domains. Information measures have been shown capable of capturing word classes by combining relevant contexts. By using them properly, we should be able to extract phoneme classes by exploiting phonotactic constraints. We can do so without explicitly having to specify the nature of these constraints.

Another benefit of information measures is that they are based on a solid mathematical foundation. Moreover, the value produced by an information measure usually is easy to interpret.

We can use mutual information to specify a metric which does not partition a lexicon. To do so we will compare phonemes' contexts rather than rely on word level comparisons. Thus the measure can employ notions of direction and locality which are pertinent to speech recognition.

2.1.4 Relevance to Lexical Access

While we are interested in the linguistic interpretation of our results, we are more directed by the needs of lexical access for continuous speech recognition. While we do not want to tie ourselves to any particular recognizer or recognition model, we do keep in mind the general kinds of search and discrimination which any recognizer must make.

To make practical use of broad phoneme classes within a recognizer, the classes must be acoustically detectable and should be robust. Presumably this means that the classes consist of phonemes which are acoustically similar and the robustness stems from avoiding making fine distinctions between them.

This study makes no use of acoustic data in forming classes. Doing so allows us to determine if the collocational constraints inherently embody acoustic similarities. It also avoids issues of spectral representation and comparison vital to acoustic distance measures.

2.2 Metric Overview

We based the measures used in our studies on information theory. Information theory is concerned primarily with the probabilistic analysis of communications systems and so seems well-suited to speech. Unlike typical information theory problems, we are not concerned with robust transmission. Instead, we are interested in the ability to predict a phoneme based on its context.

2.2.1 Metric Development

We based our metric for forming phoneme classes on *average mutual information* [12], defined as:

$$I(X; Y) = \sum_X \sum_Y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}.$$

It is the expected value of the *mutual information*, the amount of information about the event x provided by the event y . Note that the measure is reversible, making it equally correct to say that this measures the amount of information about y provided by x . An alternative description of mutual information says that it compares the probability that x and y co-occur, $P(x, y)$, against the probability of them co-occurring were they statistically independent, $P(x)P(y)$.

When $P(x, y) = 0$, we have an apparent problem, as the logarithm of 0 is undefined. We note that $\lim_{x \rightarrow 0} x \log x = 0$, and so substitute 0 for the computation in these instances.

When we use the same set for both arguments of the average mutual information, the formula reduces to:

$$I(X; X) = H(X) = - \sum_X P(x) \log_2 P(x).$$

This is known as the *entropy*. It is the expected value of the *self-information* of the event x , $-\log_2 P(x)$. The self-information can be interpreted as the number of bits of information needed to specify the event x . Thus the entropy is the mean amount of information needed to specify events in X .

We apply these measures by letting X and Y represent a set of phonemes. In

practice, the probabilities are estimated by the formula:

$$P(x) \approx \frac{|x|}{|X|}$$

where $|x|$ denotes the number of occurrences of x within some set X and $|X|$ denotes the size of the set.

To use mutual information to capture collocational constraints we choose X and Y to represent sequentially related phonemes. We use subscripts to indicate relative positions. For example, we denote the average mutual information between adjacent phonemes using the equation:

$$I(X_i; X_{i+1}) = \sum_{X_i} \sum_{X_{i+1}} P(x_i, x_{i+1}) \log_2 \frac{P(x_i, x_{i+1})}{P(x_i)P(x_{i+1})}.$$

By generalizing this we can define a mutual information measure between any sequence of phonemes and another.

Finally we want to map the phonemes into broad classes. We use $\Phi(x)$ to represent the class into which phoneme x is mapped. As an example, we can measure the average mutual information between a phoneme and the phoneme class following it by:

$$I(X_i; \Phi(X_{i+1})) = \sum_{X_i} \sum_{X_{i+1}} P(x_i, \Phi(x_{i+1})) \log_2 \frac{P(x_i, \Phi(x_{i+1}))}{P(x_i)P(\Phi(x_{i+1}))}.$$

Our measures presume there is at least one phoneme position being mapped into classes and it is from that phoneme's position that we measure relative distances. We may map additional phoneme positions as well.

There are many ways we can describe the context of a phoneme, each with a companion information measure. The simplest case, no context, corresponds to phoneme entropy. Next simplest is to use a single phoneme or broad class to the left or right. We can use a sequence of 2 (or more) phonemes or use some combination of left and right contexts. Finally, we can relax the ordering constraints and consider co-occurrence of phonemes within a window.

2.2.2 Normalization

The value of a mutual information measure is determined by the a priori probabilities of the events. We need to normalize the results in order to make them more meaningful

across different data sets. To do so we compute the *percent information extracted* by a phoneme class mapping, for example:

$$\text{PIE} = \frac{I(X_i; \Phi(X_{i+1}))}{I(X_i; X_{i+1})}.$$

This is the ratio of the information measured with the mapping to the information measured without it. At best, this measure will be 100% , since the mapping can only reduce the information available.

This PIE is different from the one used in Carter [7]. Both can be thought of as a two-step process. First we compute an information measure before and after some distortion. Then we compute the ratio of the two results. Carter's PIE is based on word entropy while all of our measures are based on phoneme comparisons.

2.3 Pronunciation Constructs

We base our studies on the pronunciation of words. We will next justify the lexicographic representation used.

2.3.1 Validity of Using Phonemes

We feel that phonemes are a reasonable unit to analyze as their psycholinguistic basis has been demonstrated. We choose phonemes over allophones because the phonemic inventory is better agreed upon.

Using phonemes implies that lexical pronunciations are sequences of segments. It is often difficult to find robust segment boundaries in an acoustic signal. Avoiding this dependence would require our using some more controversial lexical representation. We recognize this limitation, but also note that this may be surmountable for a particular recognition system by using data which captures segmentation difficulties. With it we can construct a lexicon which provides alternate word pronunciations, each with a varying number of phonemes. By comparing the analysis of this lexicon and the original, we can understand the effects of segmentation accuracy on lexical access.

2.3.2 Syllables are Too Controversial

Syllables have been shown to provide strong constraints on the organization of sounds into words [13]. For example, they constrain the basic CV structures allowed. Syllabic structure also has strong influences on acoustic realization, especially with regard to prosody and reduced syllables.

In order to make explicit use of syllable structure we need to mark that structure. How should we do this? The simplest approach would be to mark syllable boundaries, perhaps including lexical stress. A better method might parse syllables into their constituent structure.

There are two problems with these explicit markers. First, the correct syllable structure, if there is one, is not always clear. This problem manifests itself as ambisyllabic phonemes, variations in stress patterns, or even the validity of a more complex structure. Second, were we to settle on a particular syllable structure we must also choose the means of representing it. A logical approach would be to expand the phoneme symbol set to represent each phoneme and its syllabic position. This could result in a very large number of complex symbols, thus making the results difficult to interpret.

We chose to circumvent these difficulties by avoiding explicit syllable markers and instead utilizing the structure implicitly. By using the mutual information measure on a suitably small phoneme sequence, we should be able to capture many constraints imposed by syllable constituents. Longer sequence could conceivably cover long distance intrasyllable constraints such as those between the onset and coda.

As an added benefit, by not breaking the word into syllables we will be able to see cross-syllable and cross-morpheme effects. We know that many phoneme collocational constraints are not enforced at these boundaries, but this can provide much constraint by fixing the location of the boundary.

2.3.3 Word Boundary Independence

The logical unit for our studies is the word, as it is also the basis for our lexicons. Because we are interested in continuous speech recognition, we should consider more than intraword constraints; we should also examine phoneme constraints across word

boundaries. Like at syllable and morpheme boundaries, phoneme collocational constraints at word boundaries can be a powerful aid to speech recognition [14].

To include cross-word constraints we need to concatenate the phoneme sequence at the end of one word with the phoneme sequence at the start of another. An elementary approach is to do this for all words. This corresponds to using a (word)* grammar, a grammar in which a word can be followed by any other word with equal probability. While it is possible to construct a sentence containing an arbitrary sequence of words, syntax and semantics greatly constrain the set of reasonable sentences. If we wish to consider phoneme sequences across word boundaries, we should account for these word sequence constraints. However, doing so forces us to assume a particular language model.

We do not wish to rely on a language model as it would tie us to a particular task or recognition system. Accordingly, we have chosen not to examine inter-word phoneme sequences, realizing that this gives us only a partial view of the collocation constraints.

Words are clearly a unit of language, but the boundaries between words are difficult to determine in continuous speech. Therefore we should try to avoid a metric which relies on word boundaries.

We could include a word boundary marker, /#/ , in our pronunciations to explicitly capture these constraints. We think doing so is somewhat arbitrary and it ignores other word structures. Additionally, using word boundary markers may adversely affect experiments. Some lexicons are constructed to contain many regular forms of a word. This produces a preponderance of sequences like /d#/ , /z#/ , and /ŋ#/ . For both of these reasons we have decided not to represent word boundaries as a pronunciation symbol.

2.4 Clustering Technique Overview

We next describe how to combine phonemes into classes. Our basic approach is to cluster phoneme symbols based on a lexical information metric. Because we do not know the right number of classes for lexical access, we should incorporate some means of creating classes using various degrees of specificity. This suggests using some type

of hierarchical structure.

2.4.1 Many Possible Classes

There are two fundamental ways of producing a phoneme hierarchy. The first, the agglomerative approach, iteratively combines classes until a single class is formed containing all of the phonemes. The second, the divisive approach, iteratively splits classes until all classes contain a unique phoneme.

We can show that there are $2^{n-1} - 1$ ways to divide n phonemes into 2 classes. If we have on the order of 40 phonemes, the first split in forming a binary class tree would require evaluating roughly $2^{39} - 1 \approx 5 \times 10^{11}$ possibilities. This is far too many for practical purposes. Instead we chose to use a pair-wise agglomerative approach to form the hierarchy.

2.4.2 Algorithm

The algorithm for forming the hierarchy is simple. Initially, consider a set of n phoneme symbols. We try each of the $\binom{n}{2}$ possible symbol pairs for clustering. For each pair, we temporarily map the lexicon to replace occurrences of the second symbol in the pair with the first. Then we compute the PIE resulting from this map. We cluster the symbol pair which maximizes the PIE and permanently map the lexicon using it. This reduces the symbol set size to $n - 1$. We iterate the procedure until only a single symbol is left.

When there is a single symbol representing all the phonemes there is no information present to predict. Thus a necessary terminal condition of the clustering is that the PIE is zero. This contrasts with Carter, where a single-symbol mapping still retains the information present in the number of phonemes per word.

2.4.3 Reducing Computational Complexity

At first this algorithm seems computationally intensive because we must map the lexicon many times. Worse, the time needed to do so grows with the size of the lexicon. This presents a computational problem, since we need to use large lexicons so as to best approximate the actual usage.

Note that the reason we do the mapping is to change the parameters of the mutual information measure. These arguments are restricted to only a portion of each pronunciation. Thus we can collapse the lexicon to an m -dimensional table counting occurrences of m -long phoneme sequences. This table supplies values for estimating $P(x,y)$. Similarly we can construct tables for the marginals $P(x)$ and $P(y)$.

These tables alone are sufficient for calculating the mutual information measure. They also eliminate the costly lexicon mapping. We calculate the effect of merging two classes by summing the entries corresponding to those classes across other dimensions in the table. This procedure makes the algorithm independent of lexicon size except for the initial table generation. It depends only on the number of phonemes and the length of the sequences.

2.5 Lexicon Preparation

We base most of our studies on a modified version of the Merriam Webster Pocket Dictionary, which we will refer to as "MPD." The lexicon contains roughly the 20,000 most common words in American English. We chose this lexicon because many researchers have refined and checked its pronunciations. In addition, using it will allow us to compare our results to some earlier work.

2.5.1 Modifying Pronunciations

We want to be able to compare our results to those produced by distinctive features because their constraints for lexical access are poorly understood. Features, in the form of feature vectors, have difficulty representing some of the symbols used in MPD, notably the diphthongs and syllabic consonants. In addition, lexical stress in MPD is indicated both through the use of stress markers and by schwas in reduced stress syllables. To eliminate these problems we apply the following rewrite rules:

$$/\partial/ \rightarrow / \wedge / \quad / \mathbf{i}/ \rightarrow / \mathbf{i}/ \quad / \partial \cdot / \rightarrow / \mathfrak{z} /$$

$$/\mathbf{l}/ \rightarrow / \wedge \mathbf{l} / \quad / \mathbf{m}/ \rightarrow / \wedge \mathbf{m} / \quad / \mathbf{n}/ \rightarrow / \wedge \mathbf{n} / \quad / \eta / \rightarrow / \wedge \eta /$$

$$/\text{ɔ}^y/ \rightarrow / \text{ɔi} / \quad / \text{a}^y/ \rightarrow / \text{ai} / \quad / \text{a}^w/ \rightarrow / \text{au} /$$

No change is made to the diphthongs $/i^y/$, $/e^y/$, or $/o^w/$ since their monophthong counterparts, $/i/$, $/e/$, and $/o/$, can be represented by feature vectors. In addition, all syllable and stress markers are removed from the pronunciations.

2.5.2 Word Frequency Weighting

Previous studies have shown that weighting the lexicon using frequency of occurrence in the Brown Corpus [15] can have dramatic effects due to the overwhelming influence of common words. We think of such weighting as a zeroth-order language model. Using any language model opens a host of issues which we could not adequately address in this thesis. Accordingly we chose to ignore weighting effects for clustering.

2.6 An Example

We will next present an example of our clustering technique so that the reader can better understand it. For this example we will use a subset of MPD consisting of 33 words:

Spelling	Pronunciation	Spelling	Pronunciation	Spelling	Pronunciation
aisle	ail	lie	lai	peel	pil
ay	ai	lisle	lail	peep	pip
aye	ai	loll	lal	pi	pai
eel	il	lollypop	lalipap	pie	pai
eye	ai	lop	lap	pile	pail
I	ai	lye	lai	pipe	paip
isle	ail	P	pi	plea	pli
lea	li	pa	pa	plop	plap
leal	lil	pas	pa	ply	plai
leap	lip	pea	pi	pop	pap
lee	li	peal	pil	poppy	papi

We selected these words because they are the largest number of words which can be formed using only four different phonemes.

Let's find the phoneme clusters created using the mutual information between adjacent clusters as the metric. There are 65 diphones in this sub-lexicon of which only 10 are unique. We estimate the probability of a diphone occurring from its

frequency in the data and summarize the results in a table:

$$P(x_i, x_{i+1}) \left\{ \begin{array}{c|c|c|c|c} & l_i & a_i & i_i & p_i \\ \hline l_{i+1} & & \frac{2}{65} & \frac{8}{65} & \frac{3}{65} \\ \hline a_{i+1} & \frac{8}{65} & & & \frac{9}{65} \\ \hline i_{i+1} & \frac{6}{65} & \frac{14}{65} & & \frac{6}{65} \\ \hline p_{i+1} & & \frac{5}{65} & \frac{4}{65} & \\ \hline \end{array} \right\} P(x_{i+1})$$

$$\underbrace{\begin{array}{c|c|c|c} \frac{14}{65} & \frac{21}{65} & \frac{12}{65} & \frac{18}{65} \\ \hline \end{array}}_{P(x_i)}$$

Using these data we can compute the desired mutual information and PIE:

$$\begin{aligned} I(X_i; X_{i+1}) &= \sum_{X_i} \sum_{X_{i+1}} P(x_i, x_{i+1}) \log_2 \frac{P(x_i, x_{i+1})}{P(x_i)P(x_{i+1})} \\ &= 0 + \frac{2}{65} \log_2 \frac{\frac{2}{65}}{\frac{21}{65} \cdot \frac{13}{65}} + \frac{8}{65} \log_2 \frac{\frac{8}{65}}{\frac{12}{65} \cdot \frac{13}{65}} + \frac{3}{65} \log_2 \frac{\frac{3}{65}}{\frac{18}{65} \cdot \frac{13}{65}} + \\ &\quad \frac{8}{65} \log_2 \frac{\frac{8}{65}}{\frac{14}{65} \cdot \frac{17}{65}} + 0 + 0 + \frac{9}{65} \log_2 \frac{\frac{9}{65}}{\frac{18}{65} \cdot \frac{17}{65}} + \\ &\quad \frac{6}{65} \log_2 \frac{\frac{6}{65}}{\frac{14}{65} \cdot \frac{26}{65}} + \frac{14}{65} \log_2 \frac{\frac{14}{65}}{\frac{21}{65} \cdot \frac{26}{65}} + 0 + \frac{6}{65} \log_2 \frac{\frac{6}{65}}{\frac{18}{65} \cdot \frac{26}{65}} + \\ &\quad 0 + \frac{5}{65} \log_2 \frac{\frac{5}{65}}{\frac{21}{65} \cdot \frac{9}{65}} + \frac{4}{65} \log_2 \frac{\frac{4}{65}}{\frac{12}{65} \cdot \frac{9}{65}} + 0 \\ &\approx 0.719 \text{ bits} \end{aligned}$$

$$\text{PIE} = 100\%$$

Using the 4 phonemes, there are $\binom{4}{2} = 6$ ways we can form 3 classes. We try each of these possibilities and note which one has the greatest PIE:

	l_i	a_i	i_i	p_i
l_{i+1}	$\frac{10}{65}$	$\frac{8}{65}$	$\frac{12}{65}$	
a_{i+1}	$\frac{20}{65}$		$\frac{6}{65}$	
i_{i+1}	$\frac{5}{65}$	$\frac{4}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.272 \text{ bits}$$

$$\text{PIE} \approx 38\%$$

	l_i	i_i	a_i	p_i
l_{i+1}	$\frac{14}{65}$	$\frac{16}{65}$	$\frac{9}{65}$	
i_{i+1}	$\frac{8}{65}$		$\frac{9}{65}$	
a_{i+1}	$\frac{4}{65}$	$\frac{5}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.243 \text{ bits}$$

$$\text{PIE} \approx 34\%$$

	l_i	p_i	a_i	i_i
l_{i+1}	$\frac{3}{65}$	$\frac{7}{65}$	$\frac{12}{65}$	
p_{i+1}	$\frac{17}{65}$			
a_{i+1}	$\frac{12}{65}$	$\frac{14}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.610 \text{ bits}$$

$$\text{PIE} \approx 85\%$$

	a_i	i_i	l_i	p_i
a_{i+1}	$\frac{14}{65}$	$\frac{14}{65}$	$\frac{15}{65}$	
i_{i+1}	$\frac{10}{65}$		$\frac{3}{65}$	
l_{i+1}	$\frac{9}{65}$			

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.283 \text{ bits}$$

$$\text{PIE} \approx 39\%$$

	$a_i p_i$	l_i	i_i
a_{i+1}	$\frac{14}{65}$	$\frac{8}{65}$	$\frac{4}{65}$
p_{i+1}	$\frac{5}{65}$		$\frac{8}{65}$
l_{i+1}	$\frac{20}{65}$	$\frac{6}{65}$	

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.297 \text{ bits}$$

$$\text{PIE} \approx 41\%$$

	i_i	p_i	l_i	a_i
i_{i+1}	$\frac{10}{65}$	$\frac{6}{65}$	$\frac{19}{65}$	
p_{i+1}	$\frac{11}{65}$		$\frac{2}{65}$	
l_{i+1}	$\frac{9}{65}$	$\frac{8}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.363 \text{ bits}$$

$$\text{PIE} \approx 51\%$$

From this we determine it is best to merge /l/ and /p/ into a single class. We keep this merger and consider the 3 possible sets of 2 classes:

	l_i	p_i	a_i	i_i
l_{i+1}				
p_{i+1}			$\frac{27}{65}$	$\frac{12}{65}$
a_{i+1}				
i_{i+1}		$\frac{26}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.156 \text{ bits}$$

$$\text{PIE} \approx 22\%$$

	l_i	p_i	i_i	a_i
l_{i+1}				
p_{i+1}			$\frac{27}{65}$	$\frac{21}{65}$
i_{i+1}				
a_{i+1}		$\frac{17}{65}$		

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.178 \text{ bits}$$

$$\text{PIE} \approx 25\%$$

	l_i	p_i	a_i	i_i
l_{i+1}				
p_{i+1}		$\frac{3}{65}$	$\frac{19}{65}$	
a_{i+1}		$\frac{29}{65}$	$\frac{14}{65}$	
i_{i+1}				

$$I(\Phi(X_i); \Phi(X_{i+1})) \approx 0.203 \text{ bits}$$

$$\text{PIE} \approx 28\%$$

Of these possibilities, merging /a/ and /i/ produces the best results. Finally, we are left with only one merger possible:

	l_i	p_i	a_i	i_i
l_{i+1}				
p_{i+1}			$\frac{65}{65}$	
a_{i+1}				
i_{i+1}				

$$I(\Phi(X_i); \Phi(X_{i+1})) = 0 \text{ bits}$$

$$\text{PIE} = 0\%$$

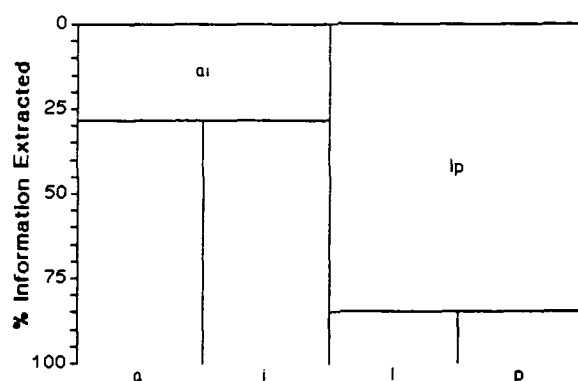


Figure 2.1: Dendrogram corresponding to the example phoneme clustering.

We can best understand the resulting hierarchy by displaying it in the form of a dendrogram [16], shown in Figure 2.1. The dendrogram's abscissa lists the phonemes in our hierarchy while the ordinate shows PIE. We denote two classes joining by drawing a horizontal line across them at the PIE level where they merged. This display allows us to see both the phoneme clusters and their relative robustness.

2.7 Summary

We have presented a technique for forming phoneme clusters using a minimum of presumed knowledge. In addition, we have chosen to use an information-theoretic metric because its utility in capturing collocational constraints has been demonstrated. We have shown how a metric can be defined over a phoneme sequence and have justified the pronunciations we use.

Chapter 3

Experiments

In this chapter we present the results of our phoneme clustering experiments. We evaluate the performance of our clusters against those suggested by other studies using both historic and new lexical metrics. Finally, we examine some of the questions related to our techniques.

Our experiments considered many variations on the basic theme. For clarity, we will present only some representative results here. Some of our additional phoneme clustering experiments are described in appendix A.

In our work we have striven to avoid the preconceived notions of how to structure phonemes; yet, we cannot conduct nor discuss our study in a vacuum. We will always need to compare our results to alternate structures, and the best frame of reference available consists of the phoneme groupings proposed by linguists, e.g., distinctive features. Accordingly, we will note similarities and differences between these two approaches whenever appropriate.

3.1 Diphones

We begin by considering the use of the minimum contextual information. In our first study, we cluster phonemes based on the average mutual information between a phoneme's class and the following phoneme's class. Because the average mutual information is a reversible measure, it does not matter whether we consider a phoneme and its successor or its predecessor. We capture constraints in both directions at once. We display the results of our clustering as a dendrogram in Figure 3.1.

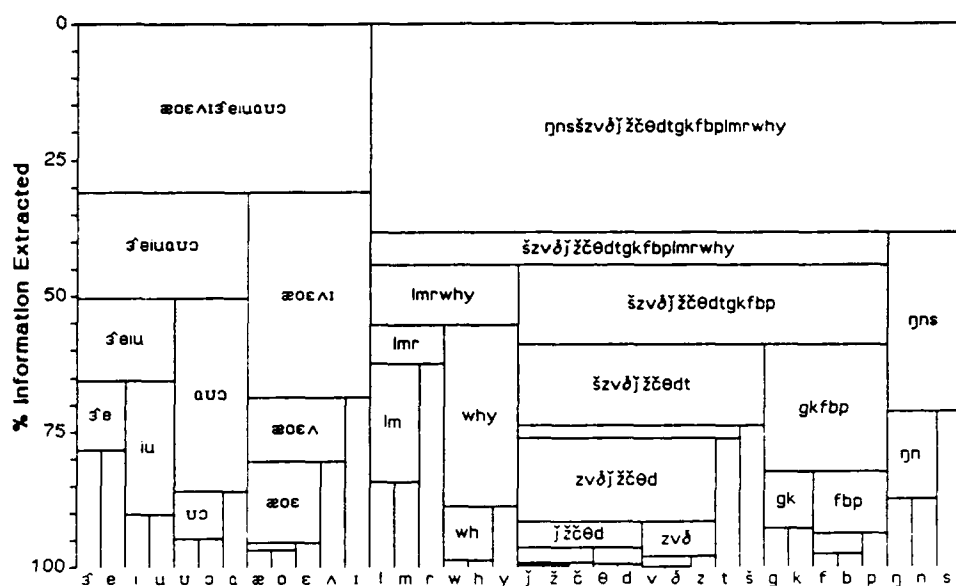


Figure 3.1: Dendrogram produced by clustering based on diphones.

Perhaps the most striking aspect of this hierarchy is that it completely segregates the vowels and consonants from one another. This single distinction provides a large amount of information, roughly 30% of what is possible.

Vowels are organized into subclasses, and some of the divisions are suggestive of dimensions used to describe vowel color. For example, /i/ and /u/ are high and tense while /ʊ/ and /ɔ/ are back and rounded.

The consonants can be viewed as being divided into four roughly-separated classes: semivowels, fricatives, stops, and nasals. The stops are divided based on their place of articulation. There may be an affinity between the coronals, demonstrated by clusters containing /l/ and /r/: /s/ and /n/: and /ʃ/, /z/, /θ/, /j/, /ʒ/, /č/, /θ/, /d/, and /t/. It is interesting to note that /h/, often an oddity for other classification schemes, is placed among the semivowels.

Many of the classes formed contain phonemes which are similar acoustically. This is fascinating, as no part of our clustering procedure requires this to be so. We suggest that this may be viewed as evidence that language's acoustic and contextual constraints may be evolving simultaneously.

The dendrogram shows that the robustness of the phoneme classes varies through-

out the hierarchy. In particular, some of the initial clusters formed are not very robust and soon merge with other clusters. Yet these clusters are crucial in determining the form of the remaining tree. Perhaps some of the phonemes which are grouped against our linguistic intuition do so because of these unstable initial steps.

We can examine these critical times in the clustering process to search for slightly lower scoring classes which give better linguistic justification. A class which is sufficiently close in score to the best might have won under different conditions, perhaps a different lexicon or a divisive search. Examination of our dendrogram confirms this suspicion. For example, even though the /fbp/ cluster is satisfactory, linguists might prefer the stops to have merged first. When we examine the clustering scores, placing /f/ with /b/ yields 97.52 PIE while /f/ with /p/ yields 97.31 PIE and /p/ with /b/ yields 97.29 PIE. These are relatively close. Thus we can attribute at least some of the "improper" clustering to competing classes which nearly yielded the most information extracted but failed to do so.

In terms of linguistic description, perhaps the worst cluster in the hierarchy is the fairly robust one combining /ɲn/ with /s/. Linguists might prefer clustering the nasals and placing the /s/ near /ʃ/ or /z/. How might we explain this clustering? In part, it may be the aforementioned gravity of the coronals. Alternatively, it may be an artifact of the lexicon having many words containing /m/, /ɱ/, or /ɿs/.

The fact that such a simple process can yield a reasonable looking dendrogram is most encouraging. Even now we can see a fair degree of agreement between linguistic theory and our data-driven approach.

3.2 Triphones

We next add additional contextual information to see if this will result in classes better fitting linguistic descriptions. We do so by considering sequences of three phonemes rather than two. Note that these metrics incorporate directionality, unlike the metric used in the previous experiment. It arises because we treat the phonemes unequally by pairing two as the "context" for the third. Thus, there are three possible ways to compute the average mutual information over a sequence of three phonemes, $x_1x_2x_3$:

$$I(X_1:(X_2, X_3)) \quad I(X_2:(X_1, X_3)) \quad I(X_3:(X_1, X_2))$$

3.3 Cluster Evaluation

We can use the hierarchies we created to partition phonemes into classes suitable for lexical access. We can use any set of classes from the hierarchy provided they encompass all of the phonemes and are mutually exclusive. Having done this we should ask how effective these classes are from a lexical access standpoint and compare them to classes suggested by other approaches.

3.3.1 Lexical Experiments

One way to evaluate our results is the type of procedure first used by Shipman and Zue [4]. They mapped a lexicon's pronunciations in accordance with six manner classes and gathered the words into cohorts. They determined the efficacy of the classes by computing a statistic over the cohorts.

Unlike previous lexical experiments which relied on a fixed number of broad classes, we can vary the number of classes used. There is a simple way to select n classes from a dendrogram for $1 \leq n \leq \text{number of phonemes}$: we "slice" the dendrogram horizontally at the level which provides the number of classes we desire. This allows us to study the tradeoff between the number of classes used and the fineness of phonetic distinctions made.

As we mentioned earlier, using word-level measures partitions the lexicon based on the number of phonemes per word. A measure which did not force this partitioning would be more appropriate for both isolated and continuous speech recognition. The information measures we used for clustering fulfill this requirement. Given a set of classes we can map the lexicon and compute the resulting PIE. By varying the underlying information measure, we can tune the lexical measure to suit a particular lexical access task.

While more abstract than cohort-based lexical measures, we think this metric is still relevant as we measure the additional information needed to distinguish a phoneme from its classmates. Doing so is one way of viewing the lexical access problem. In addition, our measure uses a logarithmic scale, which Carter [7] suggests is more appropriate for measuring the work needed to complete the lexical access task. Like Carter's metric, we can produce either a normalized result for easier comparison

on a particular task or an absolute result for comparison across tasks.

In our evaluation we provide both cohort-based and phoneme-based measures of cluster performance. The first measure we use is expected cohort size as it is representative of the cohort measures and seems more appropriate than mean cohort size. The other measure we use is the average mutual information between a class and its neighboring classes given in PIE form. This is precisely the second measurement we used for forming phoneme classes. For brevity, we refer to this measure as "context PIE." The values of some additional lexical measures are given in appendix B.

3.3.2 Baseline Establishment

Previous experiments have shown how well a set of broad classes can disambiguate a word in a lexicon. However, these experiments rely on the conviction of the reader to evaluate both the results and the metric. We would prefer an objective baseline against which a classification scheme can be measured.

One possibility is to compare the broad classification to the finest classes possible, the phonemes. This does not work well because we know it is possible to disambiguate virtually all of the lexicon using phonemes. It is similarly unreasonable to use no classes, to use a phoneme placeholder, though this is an important limit for cohort-based measures.

There is a simple way to create a baseline for broad class evaluation. We generate a hierarchy by combining phonemes at random. A typical dendrogram created this way is shown in Figure 3.3. It is important to note that this dendrogram does not embody the class structuring typical of dendrograms created using collocational constraints. We create 1000 hierarchies in this manner and compute our lexical performance measures using their classes. We then average the results to form the baseline performance.

3.3.3 Distinctive Features

We would like to compare our results to the performance of distinctive features. We use a feature set, shown in Table 3.1, based on Stevens [17]. The primary change we have made to the feature set is to specify all features' values for all phonemes. Where

[illegible]

Caption for Table 3.1: Feature-bundle specifications for phonemes used in our experiments.

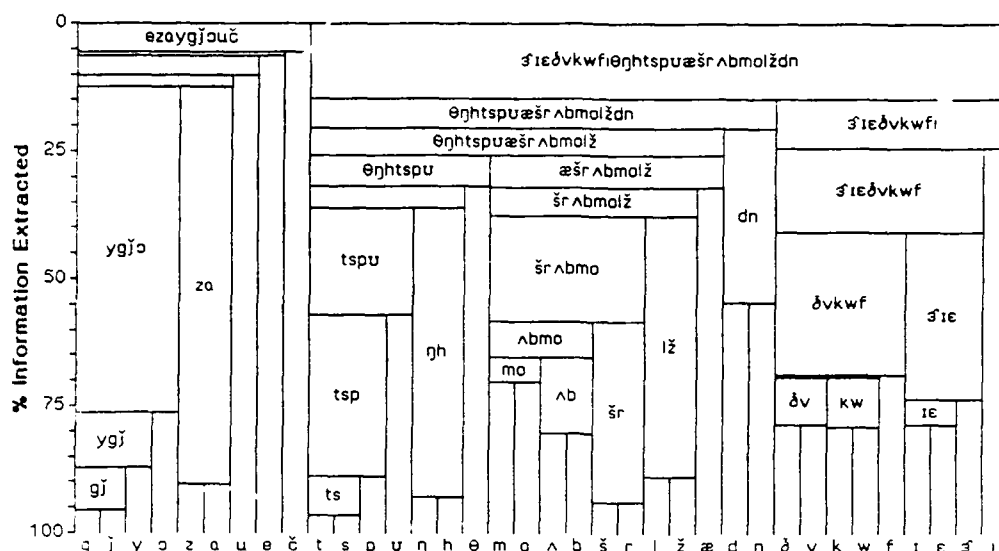


Figure 3.3: A phoneme hierarchy produced by random clustering.

Stevens left a feature unspecified, we use a “~” to indicate that the feature is not present. We have also replaced the features [Spread Glottis] and [Slack Vocal Cords] with the more common [Voiced].

Note that some of the features are unnecessary or redundant for our phoneme set. Were we to eliminate the [Nasal] feature, there would be no ambiguity amongst the phonemes. Other features, for example [Voiced], are critical in that their elimination would cause ambiguity.

By underspecifying feature values we can form phoneme equivalence classes. In order to vary the scale of the equivalence classes formed, we vary the number of features left unspecified. We begin by ranking the features to select the best single feature to use. We define best as giving the greatest cluster context PIE. We keep the phoneme distinctions made by this feature and repeat the process to select an additional feature. We iterate until all features are enabled.

This process is illustrated in Figure 3.4. The height of each bar shows the PIE for a particular subset of features. The numbered bottom axis measures the size of each feature subset. The remaining axis is divided into categories representing the features. If we looked from above we would see a triangular portion of the base covered. This is because once we enable a feature it is never disabled and we have arranged the

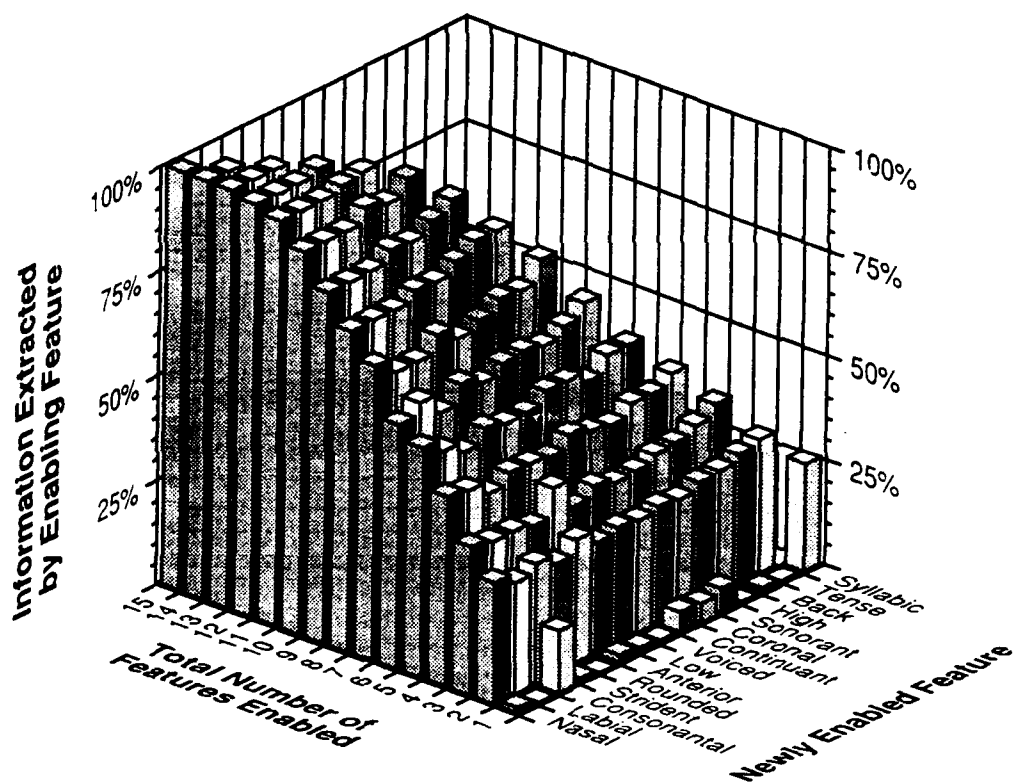


Figure 3.4: Percent of context cluster information extracted by compounded feature specification.

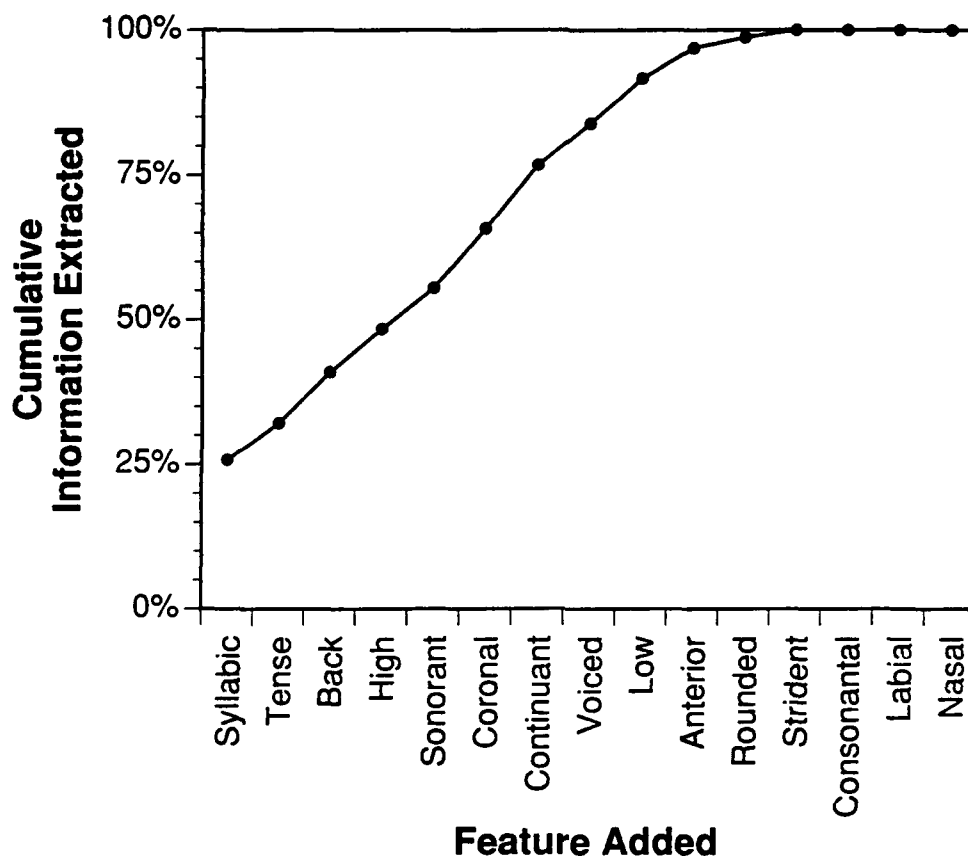


Figure 3.5: Maximum information extracted by compounded feature specification.

features in the order of their inclusion.

This figure permits us to see how the relative importance of a feature depends on those already specified. For example, by examining the row corresponding to using a single feature, we find that [Syllabic] is the best with [Consonantal] second best. The two features provide redundant information. As shown in the second row, once we have specified [Syllabic] the importance of [Consonantal] diminishes. Instead, the [Tense] feature, insignificant in the one-feature ranking, is now the best one to add. Although it was relatively important initially, [Consonantal] will end up providing no useful information at all.

This procedure forms a ranking of the features based on the phoneme identification information they supply. Although the information is represented in the rear diagonal of the previous figure, we reproduce it in Figure 3.5 for clarity. Although we can use n

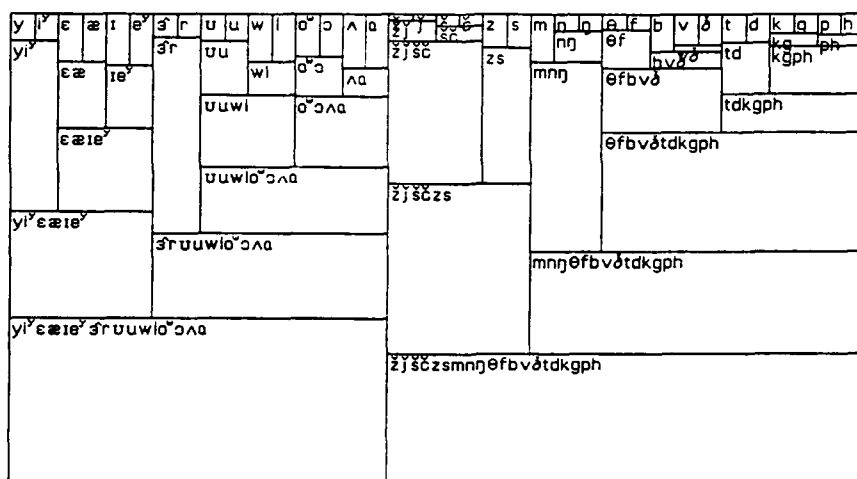


Figure 3.6: Dendrogram of a phoneme hierarchy produced using an acoustic similarity metric.

features to specify 2^n classes, the redundancies between features as well as unrealized feature bundles mean that in general fewer classes will be available.

3.3.4 Acoustic Clustering

We also compare our results to a phoneme hierarchy produced using acoustic data. For this we use a method developed by Glass [18]. This technique uses a spectral average to represent each cluster. The most similar clusters are merged in an agglomerative clustering procedure.

For our experiment, we produced a hierarchy based on phonetic transcription segments from 1000 TIMIT database [19] utterances. Rather than using all phonetic transcription characters, we reduce the set to those symbols used in our other studies. We include the diphthongized vowels where a monophthong is not used in transcribing the data.

The resulting dendrogram is shown in Figure 3.6. We can extract phoneme classes from this dendrogram using the same procedure as for our other dendrograms.

Many of the classes in this hierarchy are similar to ones produced using phoneme collocational data even though our experiments do not use any form of acoustic similarity measure. A key difference between the results is in the two cluster split:

semivowels are more like vowels than consonants in this acoustic classification, but they behave more like consonants linguistically

3.3.5 Results

We conducted our evaluation on the MPD lexicon, prepared exactly as for our phoneme clustering experiments. The results are shown in Figure 3.7.

A lower expected cohort size is interpreted as better for lexical access. The smaller the expected cohort size, the fewer average number of words we need to distinguish, and so the less we need to rely on making accurate fine phonetic distinctions.

The performance of acoustic clusters is generally worse than that of the other cases, however it is somewhat better when using only a few classes. The acoustic class performance is punctuated by discontinuities. Each of these may correspond to particularly important phoneme distinctions.

Distinctive features are narrowly the worst choice when using fewer than 10 classes. This is the range most useful for lexical access with broad classification. Note that the expected cohort size for features drops faster than for the acoustic classes.

Our collocational constraint-based clusters perform better than either of the aforementioned schemes except for the 2- and 3-class cases and performs comparably to the six manner classes. The discontinuity between 8 and 9 context clusters corresponds to the first fully identified phoneme, /r/, splintering from its parent class.

Unfortunately, using expected cohort size we find that our randomly formed clusters perform best, except when using fewer than 6 classes. Since we believe phoneme classes derived using speech knowledge should perform better than those derived randomly, we conclude that expected cohort size is a poor measure of lexical access difficulty.

We expect a larger value to signify better classes when using context PIE. Larger values mean we have come closer to identifying phonemes in the lexicon. Presumably this means we are close to identifying words as well.

We first check the performance of randomly-formed classes. These classes now clearly perform poorly compared to the others. Based on this alone, we have reason to believe this metric is superior to expected cohort size.

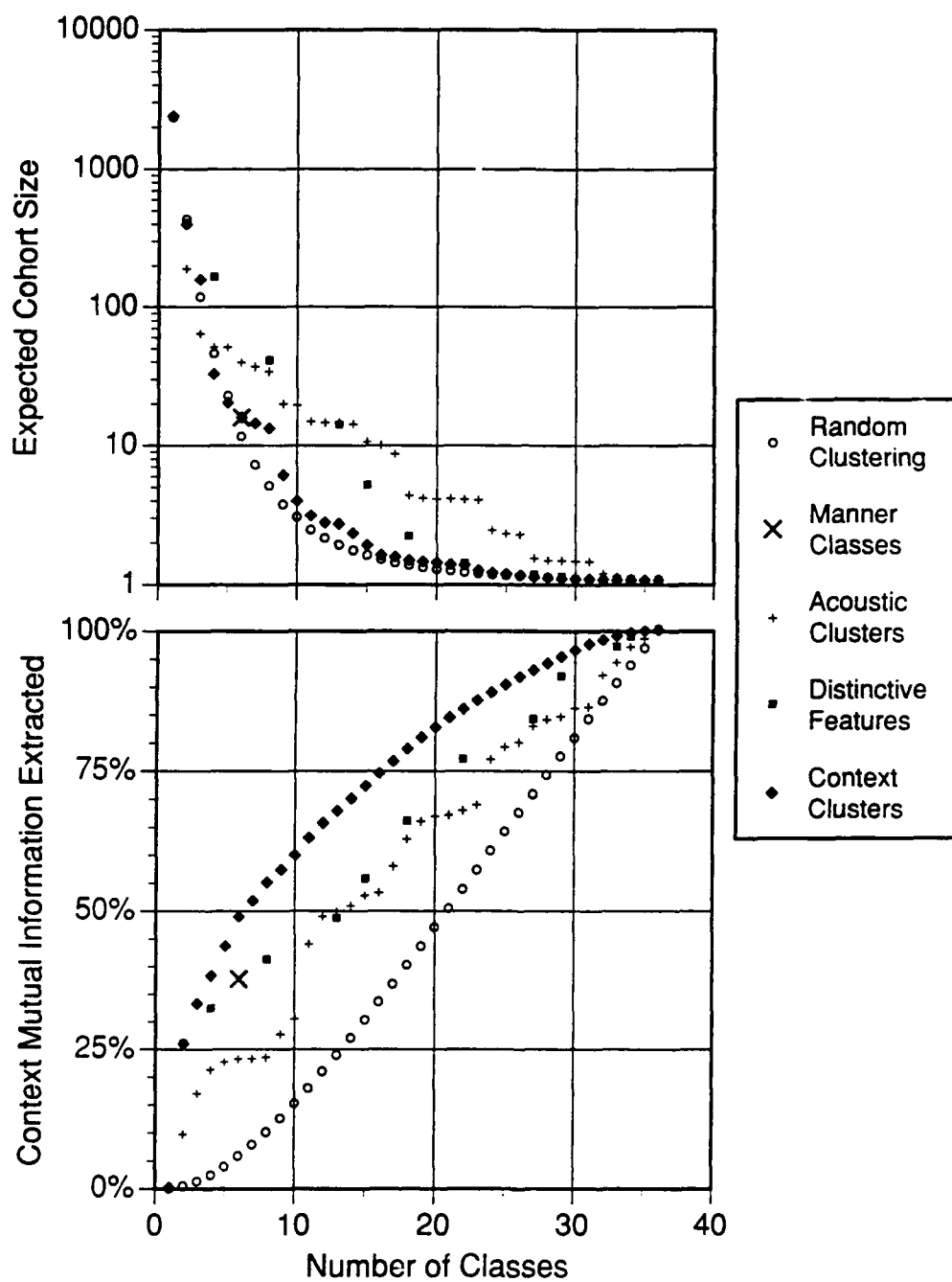


Figure 3.7: Graphs of phoneme class performance as measured by expected cohort size (top) and context PIE (bottom).

Classes based on context PIE perform best regardless of how many categories are used. This is not surprising since the clustering procedure finds locally optimum classes for this mutual information measure. Were we to expand our search to seek the optimum set, we could provide an upper bound on class performance using this metric.

Using this measure we find features perform relatively well for less than ten classes. The manner classes perform on par with distinctive features. The acoustic classes perform notably worse over this range, but later move to match the performance of features.

The two class case is particularly interesting. Feature- and context-based clusters yield just over 25% of the lexicon's information while acoustically-based clusters yield only about 10%. The only difference in the class pairs they use is that acoustic clusters place the semivowels with the vowels instead of with the consonants. The difference in performance must stem from this.

3.4 Discussion

Our study raises a number of questions we shall proceed to address. We shall take a closer look at the phoneme collocational data and examine how some of the decisions we made affect our results.

3.4.1 Capturing Collocational Constraints

How can we be sure the dendrograms we produced look the way they do because of phoneme collocational constraints? For assurance we perform a similar experiment in which we use no contextual information. We repeat our clustering procedure using PIE based on phoneme class entropy. Another way of viewing this is we base our clusters on phoneme frequency of occurrence alone. The resulting dendrogram is shown in Figure 3.8.

This dendrogram is very much unlike the dendrograms we see in our other experiments. The clusters formed do not stand out as anything linguistically relevant with two possible exceptions: /n/ with /t/ (which have the same place of articulation) and the neighborhood of /ž/. Perhaps the most striking difference is that the consonants

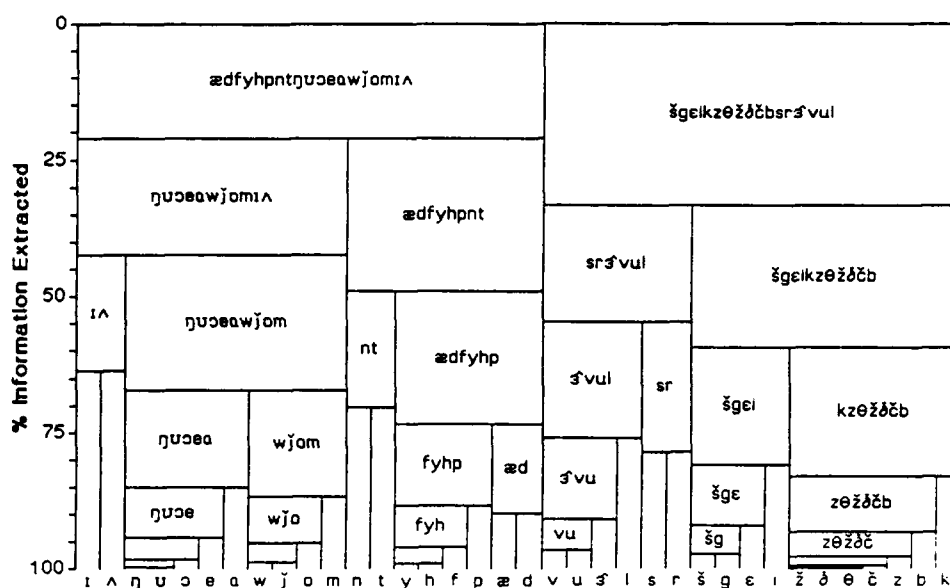


Figure 3.8: Dendrogram produced by clustering based class entropy.

and vowels are intermingled throughout this dendrogram.

This suggests that our other experiments are exploiting phoneme collocational constraints and that these constraints are fairly powerful.

3.4.2 Relationship to Pattern Recognition

How can we gain a better understanding of the information captured by our metrics? An alternative view of our hierarchical clustering procedure is that it gathers phonemes which occur in similar contexts. In the dendrogram of Figure 3.1, phoneme classes are combined when the distributions of their adjacent classes are similar. Although we have presented the experiments from an information-theoretic standpoint, we can also view the work as a pattern classification problem.

By examining these distributions we can gain a better understanding of the clustering decision process. For example, in Figures 3.9 and 3.10 we show the distribution of all phonemes following /b/ and /p/, and /ɛ/ and /ɪ/, respectively. The data is normalized so that the area under the curves is equal. Each figure shows two phonemes which are similar in terms of manner and place.

It is important to note both the similarities and differences in these figures. Both

of the stops share a profile dominated by vowels and the consonants /l/, /y/, /r/, and /s/. The phonemes following /ɛ/ and /ɪ/ show a radically different distribution from those following /b/ and /p/.

Our diphone mutual information measure used the distributions for both preceding and succeeding classes. These distributions can differ greatly, as illustrated in Figures 3.11 and 3.12. The effects of the homorganic nasal-stop rule are clearly indicated in the first distribution. In fact, the phonemes following /ŋ/ are dominated strongly by /k/ and /g/. Thus looking forward there seems to be little similarity between the nasals. Looking backwards provides more similar profiles, and again /ŋ/ seems better constrained.

Using such data we can see how phonemes with similar contexts cluster together. Note that our mutual information measure is based strictly on these distributional constraints, yet many of the clusters formed are also acoustically similar.

3.4.3 Effects of Altering Pronunciations

How did changing the pronunciations of the MPD lexicon affect our clustering? To answer this we repeated the diphone experiment using the unadulterated lexicon. We included all of the symbols, even the stress and syllable markers. The results are shown in Figure 3.13.

Here we see numerous effects not present in the original experiment. The syllable markers, /</, /-/, and /*/, are clustered together as are the stress markers /' / and /`/. The nuclei of reduced syllables, all three varieties of schwa as well as the syllabic consonants, form a cluster.

The differing symbol sets makes it difficult to compare the phoneme clusters to those of the altered pronunciations. We still see similarities between the two, particularly at higher levels in the dendrogram. We also note that some of the initial clusters formed are less robust than for the mapped lexicon.

3.4.4 Lexicon Idiosyncrasies

How are our results affected by the limitations, of both size and structure, inherent in the lexicon? A lexicon is produced within a set of guidelines to ensure pronunciation

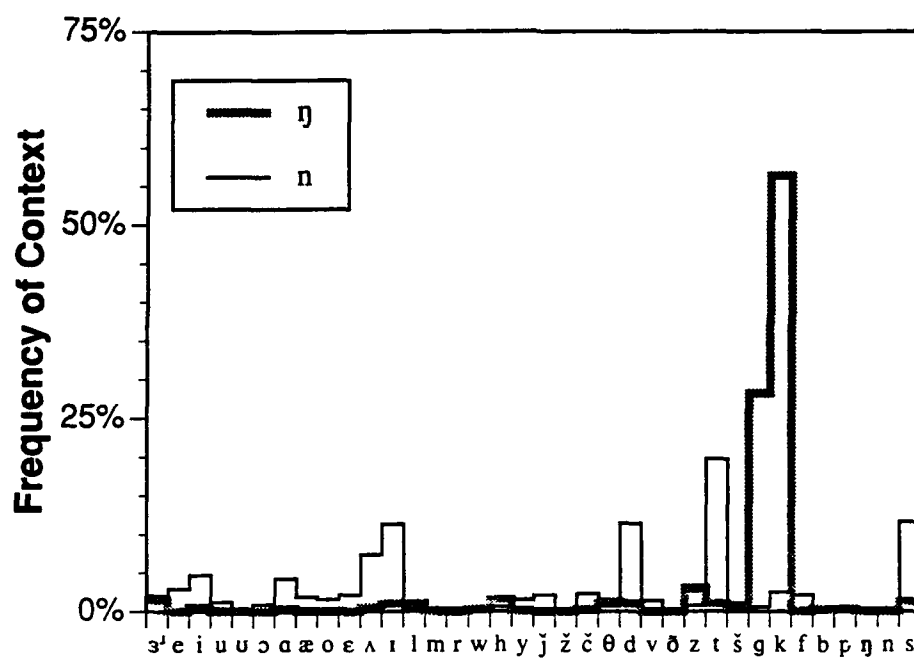


Figure 3.11: Relative frequencies for phonemes following two nasals.

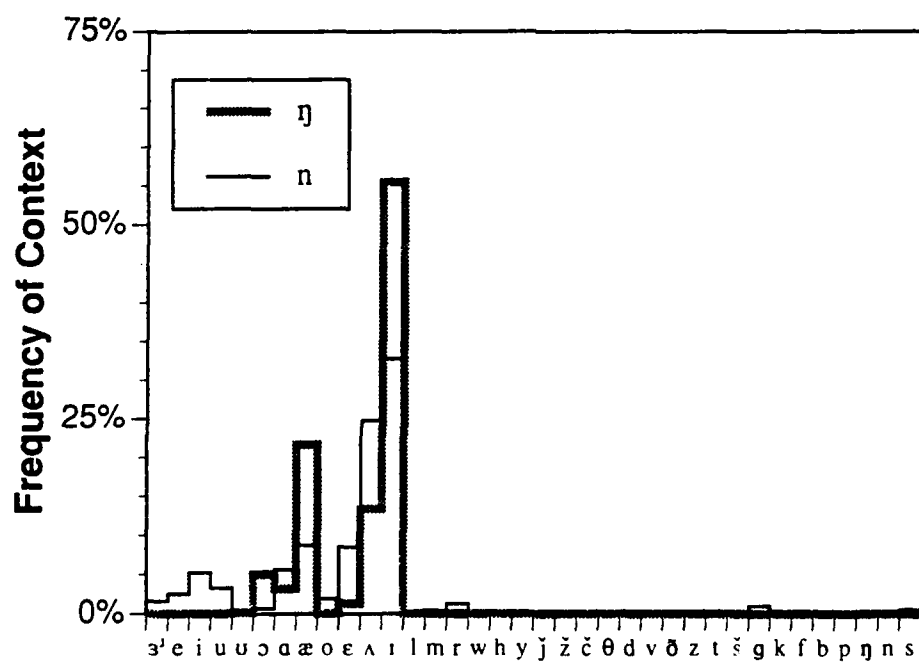


Figure 3.12: Relative frequencies for phonemes preceding two nasals.

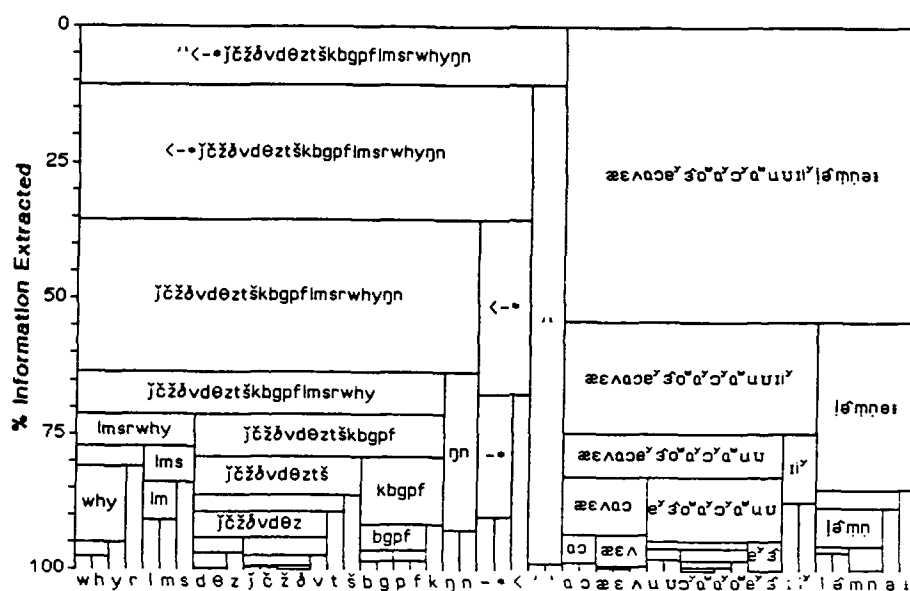


Figure 3.13: Dendrogram produced by clustering based on diphones.

consistency. Our process cannot distinguish between these rules and true linguistic constraints. Also, our results may improve if we use a larger lexicon as it will provide more stable initial clusters.

To understand these effects we repeated our trigram experiment on two additional lexicons. The first is the Shoup lexicon [20]. It uses cover symbols which represent a set of phonemes in an attempt to capture phonemic variability. The cover symbols work well when there is at most one in a word's pronunciation. When multiple cover symbols are present it may be overly generous. It includes many regular forms. The second lexicon is the MobyPronunciator lexicon [21]. It allows multiple pronunciations per word but lists them explicitly. It includes only irregular word forms. The Moby lexicon contains foreign terms which we have removed based on the use of non-English characters in their orthography.

We processed both lexicons in a manner similar to our altering of MPD. However, we permitted the Shoup lexicon to retain monophthongs not found in the others. Summary statistics for these three lexicons are shown in Table 3.2.

The results of clustering based on these lexicons are shown in Figures 3.14 and 3.15. The Shoup lexicon produces a dendrogram which hints at the MPD results but

	MPD	Moby	Shoup
Number of Pronunciations	19,837	161,675	494,569
Maximum Pronunciation Length	17	30	24
Mean Pronunciation Length	6.52	8.61	8.67
Median Pronunciation Length	6	8	8

Table 3.2: Summary statistics for three large lexicons.

is clearly not as good. We suspect this is an artifact of the over generation present in the lexicon. The Moby lexicon produces a hierarchy similar to the MPD results. Again, the clusters higher in the dendrogram have more in common than those at lower levels.

3.5 Summary

In this chapter we have shown how we generate phoneme clusters using a mutual-information metric. The classes are formed of phonemes that often share linguistic properties. We have presented the same metric as a new way of measuring the power of a set of classes within a broad-classification lexical access scheme. We have shown how we can provide both upper and lower bounds for comparing existing classification schemes on this scale. Finally we have examined some of the factors which affect our results.

Chapter 4

Conclusions

Phoneme collocational constraints provide a fertile and largely unexplored area of research. This thesis cannot possibly address all aspects of this vast subject. In this chapter we will summarize our work. We will conclude by offering a few possible extensions of our work and describe how they might be accomplished.

4.1 Summary of Results

Previous studies have shown the utility of a set of broad phoneme classes for lexical access. Broad classes can significantly constrain word candidates from a lexicon. They can avoid fine acoustic distinctions and so may be detected more robustly. A speech recognition system can exploit these two features to improve performance while reducing computation.

We demonstrated how phoneme collocational constraints can be applied to produce a hierarchy of phoneme classes. We use mutual information as a metric because of its success in capturing word collocational constraints. The classes we construct are reminiscent of linguistic and acoustic classes even though we did not tap these knowledge sources. This can be viewed as evidence of a global constraint optimization affecting phonological, lexical, and acoustic domains.

We repeated the lexical studies of previous experiments to demonstrate the constraining power of our phoneme classes. Because we arrange the phonemes into a hierarchy, we can evaluate classes of varying coarseness. In these studies our results compare favorably to those of other phoneme classification strategies.

We also have shown that the lexical metric used in these studies ranks a baseline of random classes as being better than other techniques. This spurred us to apply an alternate lexical metric based on phoneme collocational data. This metric ranks random classes below others, as a good metric should.

4.2 Suggested Extensions

There are many ways to enhance the work we have presented. Rather than provide an exhaustive list, we will consider only those we believe are most important.

4.2.1 Word Sequence Modelling

Researchers conducted earlier lexical studies at a time when large vocabulary continuous speech recognition was impractical. Accordingly, these studies were geared for isolated word systems. The questions of lexical access complexity are still relevant, but we must shift our focus to more natural speech.

We did not address across-word phoneme constraints because we did not wish to introduce a language model variable into an already complex study. Adding the necessary data is a relatively easy task. If the phoneme sequences we are interested in are sufficiently shorter than the words, a bigram language model should be adequate. Longer phoneme sequences would require more complex language models to ensure accuracy.

The results of such studies would reflect the continuous speech lexical access task more accurately than does our current work.

4.2.2 Significance Pruning of Seed Sequences

Some of the initial clusters in our dendrograms are not robust. The initial clusters are important when we are using agglomerative hierarchy construction. Because these classes make the finest distinctions, they are formed when data are most likely to be sparse. We propose using a significance test to prune entries from our sequence frequency table before clustering as a way of improving our classes.

4.2.3 Alternate Clustering Techniques

We have chosen to explore a single clustering technique using a single type of metric. Our choices were motivated by previous studies. We have shown how phoneme cluster generation can be viewed as a pattern classification problem. We can apply other pattern classification techniques for exploring phoneme collocational constraints to determine which is best. We can also compare the results of other distance metrics, for example Euclidean distance between phoneme frequency contours.

4.2.4 Recognizer Tuned Clusters

We can adapt our procedures to the performance of a particular speech recognizer. We use the recognizer's lexicon and language model. We can map the lexicon using the recognizer's confusion matrix and insertion/deletion statistics to simulate the input to the lexical access component better. This should give us a more realistic estimate of broad classification power.

4.2.5 Acoustic Detectability

Broad phoneme classes formed by our procedure are of little use to a recognizer if we cannot detect them reliably. We could incorporate acoustic distance measures for the classes into our clustering procedure, but this is not what we really seek. Instead, we propose building acoustic detectors for our broad classes. There are established means for evaluating the performance of these classifiers. Should we discover a particular class cannot be detected reliably, we can inhibit its formation in the classification tree.

There is an alternative way we can evaluate our classes based on acoustic data. By summing rows and columns in a particular speech recognizer's confusion matrix, we can approximate how our broad classes will affect that system's performance. We can use the entropy of the matrices to compare phoneme classification schemes.

4.2.6 Measuring Tree Stability

We have used lexical measures to evaluate our phoneme classes objectively. We have not provided an evaluation of the phoneme hierarchy itself. There may be ways we can

compare two hierarchies, perhaps based on the two halves of a lexicon, to determine the stability of the classes.

4.3 Summary

This thesis provides an approach for evaluating phoneme collocational constraints as they apply to lexical access. We have shown results which demonstrate the powerfulness of these constraints. We also have exposed a potential problem with earlier lexical studies on broad phoneme classification. Finally, we offer support for linguistic theories of phoneme structuring.

Related Clustering Experiments

In this appendix we present some additional clustering experiments based on phoneme collocational constraints. We will briefly describe each experiment, show the resulting dendrogram, and discuss the outcome.

A.1 Directional Diphone Measure

Although the mutual information measure is reversible, we may create a directional measure using it by treating its arguments unequally. One way to do this for diphones is to use a phoneme’s class and the adjacent phoneme as arguments rather than use two classes. Thus we can allow a class to predict the following phoneme using the measure $I(\Phi(X_i); X_{i+1})$. An alternative interpretation is that each phoneme predicts the preceding class. By using $I(X_i; \Phi(X_{i+1}))$ we reverse the direction of the measure. We show the results of both in Figure A.1.

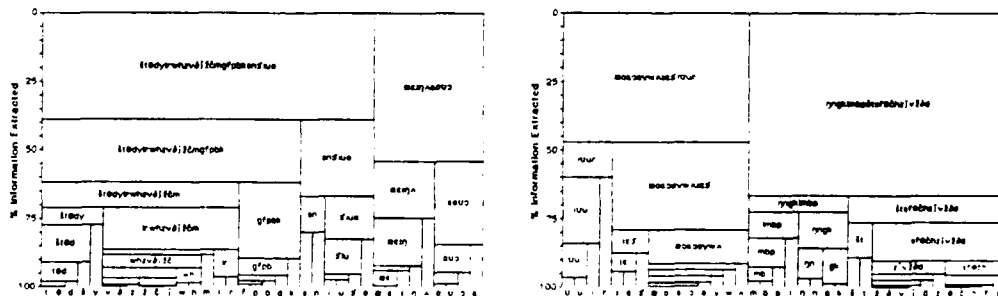


Figure A.1: Dendrograms produced by clustering based on a class predicting the following phoneme (left) and the preceding phoneme (right).

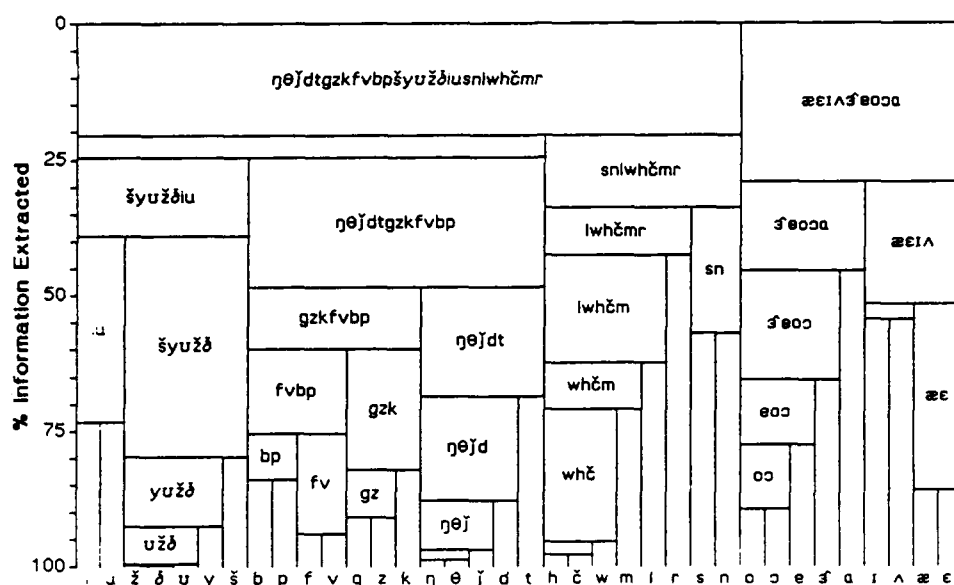


Figure A.2: Dendrogram produced by clustering based on a class predicted by the preceding two classes.

Neither of these hierarchies is as good as the one shown in Figure 3.1, especially since there isn't a clean split between the vowels and consonants. Both dendrograms have many clusters based on phonemes with similar manner or place.

A.2 Forward Prediction Triphones

Many speech recognizers use a left-to-right control strategy. A lexical access strategy based on a phoneme and its immediate context would need to process input one segment behind the acoustic decoder. We would like to determine if this delay is necessary or if we could use past context only without losing constraint. To better understand this, we used the measure $I(\Phi(X_3);(\Phi(X_1),\Phi(X_2)))$ to construct the dendrogram shown in Figure A.2.

Again, we consider the intermingling of vowels and consonants an indication that using a phoneme and its neighbors provides superior performance.

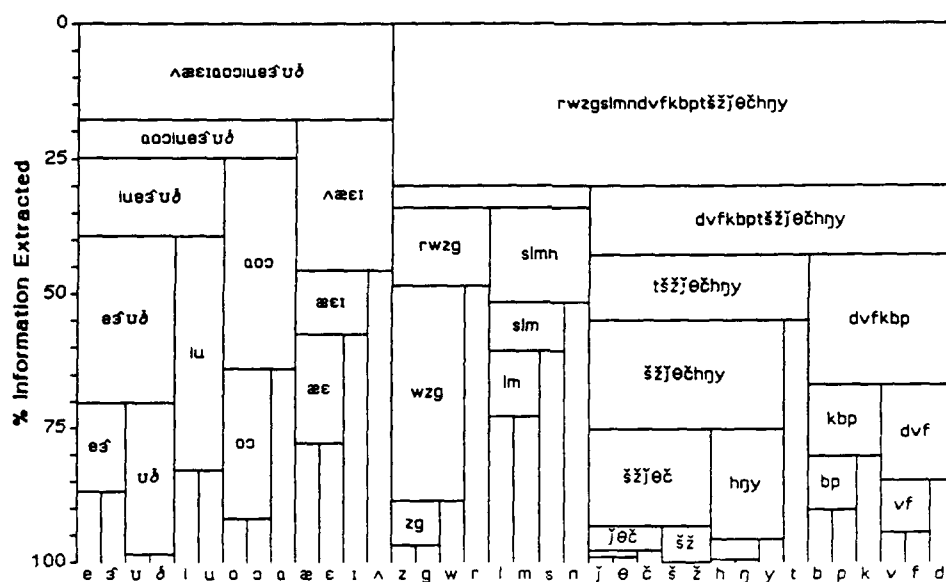


Figure A.4: Dendrogram produced by clustering based on phoneme in context using word frequency weighting.

frequent phonemes, and phonemes they are neighbors with, to behave differently.

A.5 Longer Range Effects

The dendrogram produced using phoneme in context was better than the one produced using only a single neighbor. We want to explore what will happen when we use an even larger context. As we expand the context to include more phonemes it becomes ever more important to use a language model to provide across-word sequences. We do not do this. We do, however, relax the sequence constraint in hopes of capturing more relevant events.

For these experiments, we consider the occurrence of a phoneme class in a window preceding a selected class. Thus we use a window of length $n - 1$ for a sequence of n phonemes. We show results for windows of length 2 through 5 in Figure A.5.

It is interesting to compare dendrogram (a) in this figure to Figure A.2. These differ only in the enforcement of a sequence constraint. This constraint seems to provide much information as demonstrated by the division of vowels and consonants. Also, the window approach yields a dendrogram with less robust fine clusters.

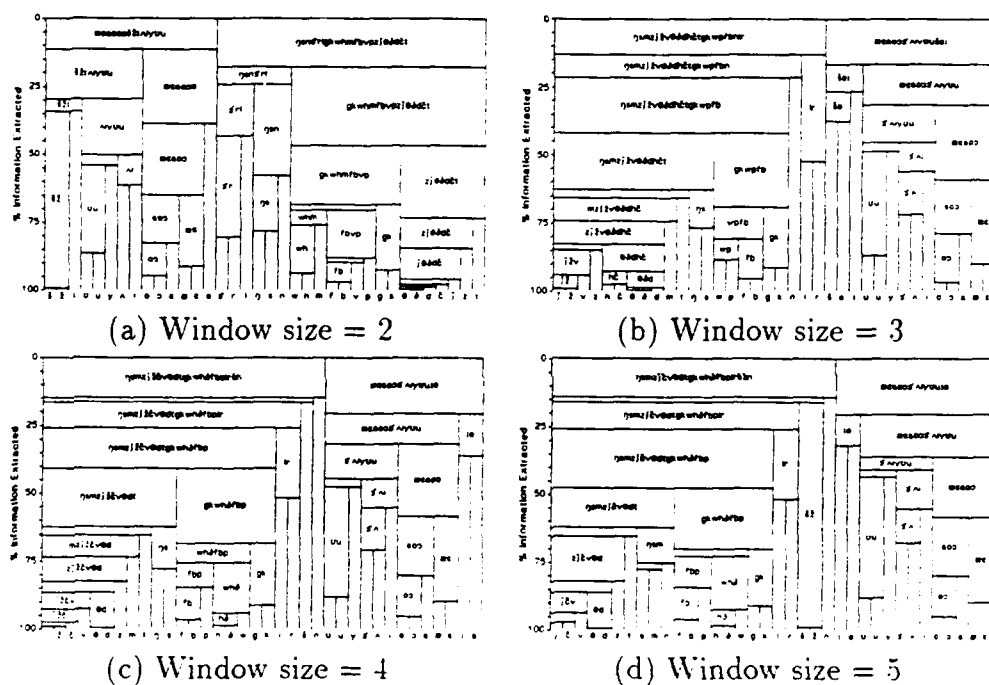


Figure A.5: Dendrograms produced by clustering based on phoneme co-occurrence within windows of varying length.

Using a longer window produces a better dendrogram, as expected. We believe the reason we don't see continued dramatic improvement for even longer windows is that we are approaching the length limit of the words.

Appendix B

Related Lexical Experiments

In this section we give the results for four additional lexical measures of broad classification performance. We do so because these measures are common in the literature. We conducted all of these experiments using the MPD lexicon.

The first measure is the percentage of words in the lexicon uniquely specified by its class pattern, shown in Figure B.1. This is considered important because it represents words which require no further acoustic discrimination for completing lexical access.

All of the classes perform similarly except for distinctive features, which performs decidedly worse. Randomly formed clusters generally perform best..

In our next graph, shown in Figure B.2, we compare classes using the word entropy measure advocated by Carter [7]. Again we see a poor separation between classification schemes, notably random classes.

We show the maximum cohort size in Figure B.3. This provides a measure of the most difficult word identification problem remaining after broad classification. Again we see the classes performing similarly except for the worse performance of features. Here we also see many discontinuities. These arise mainly from individual phonemes being identified.

Finally, we examine mean cohort size. It is a cousin of expected cohort size, but perhaps gives a less accurate picture of lexical access difficulty. The results for the measure are shown in Figure B.4.

The highly skewed nature of the data means it is difficult to make detailed comparisons. The results are similar in nature to the previous three.

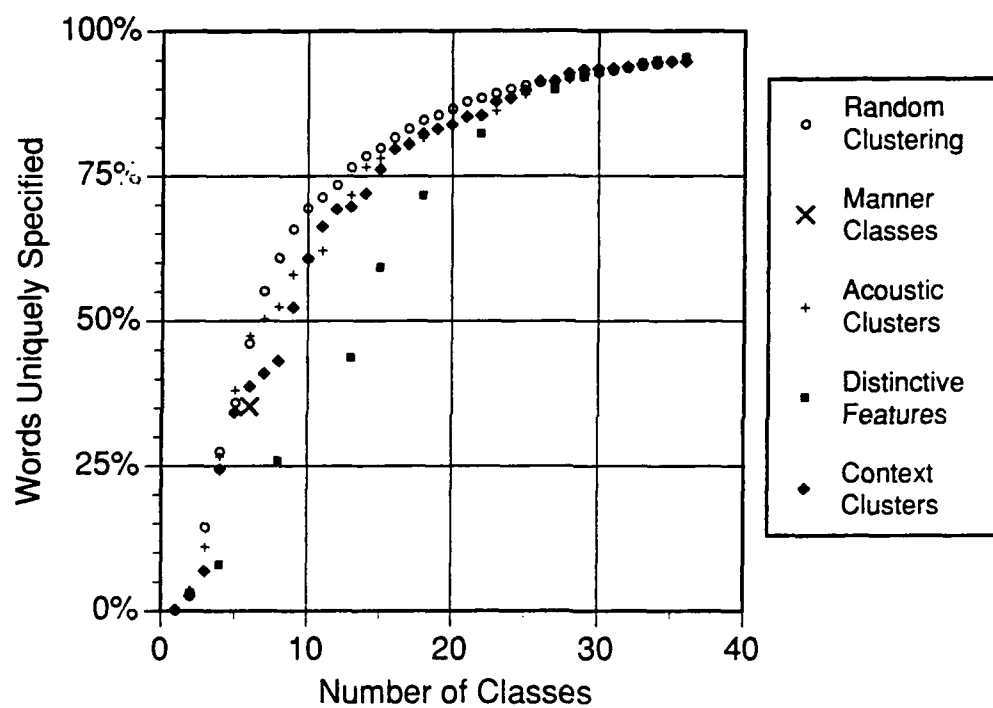


Figure B.1: Graph of phoneme class performance as measured by percentage of words uniquely specified by their class pattern.

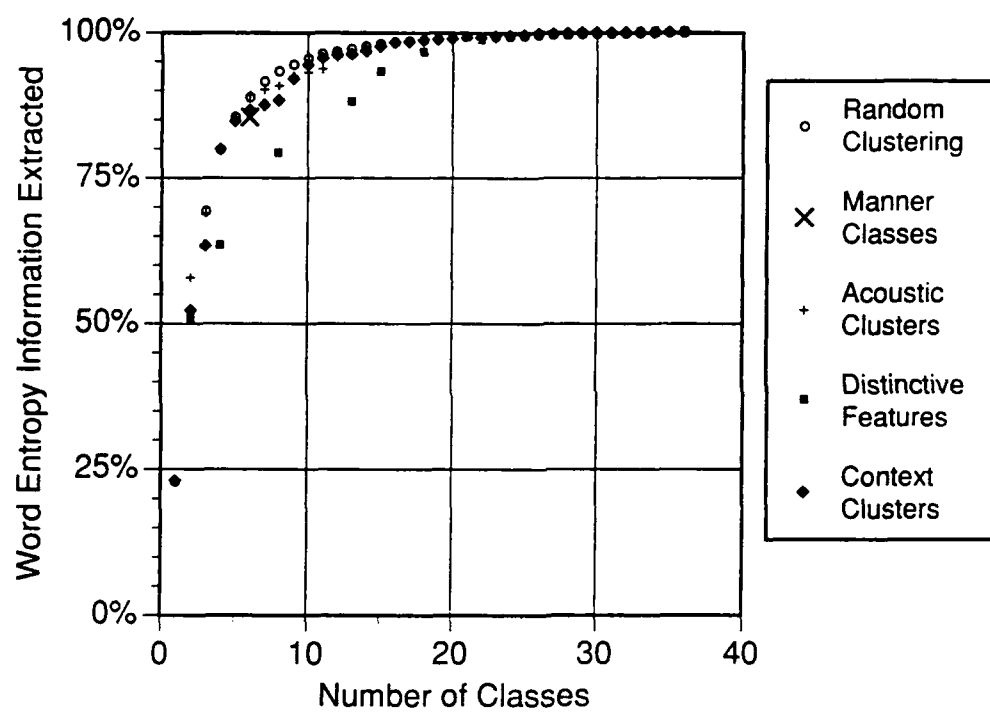


Figure B.2: Graph of phoneme class performance as measured by word entropy.

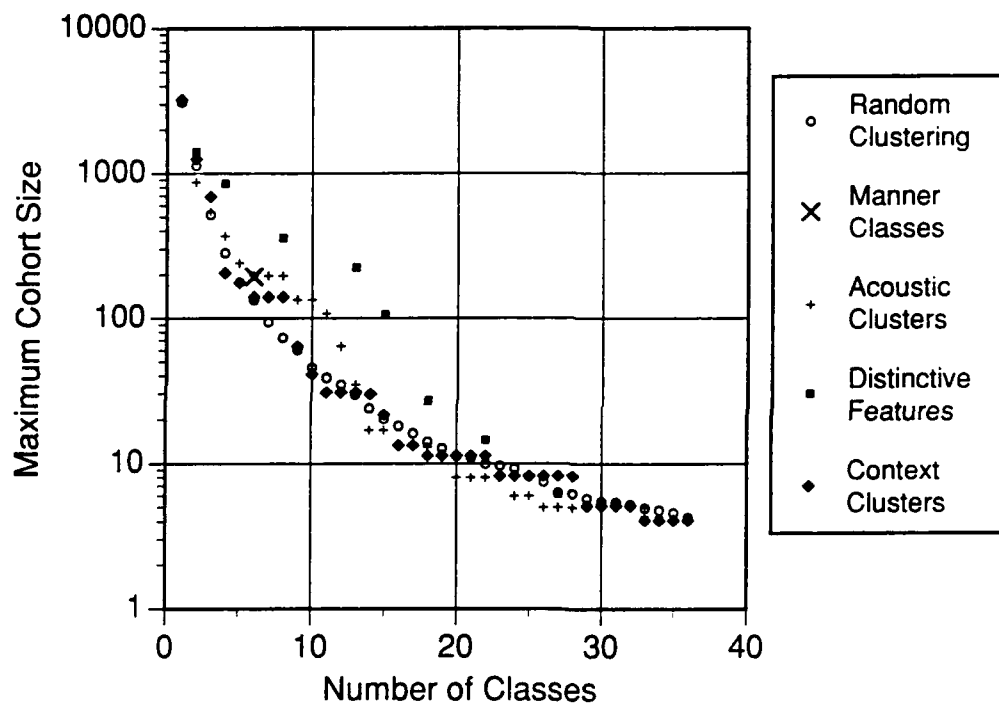


Figure B.3: Graph of phoneme class performance as measured by maximum cohort size.

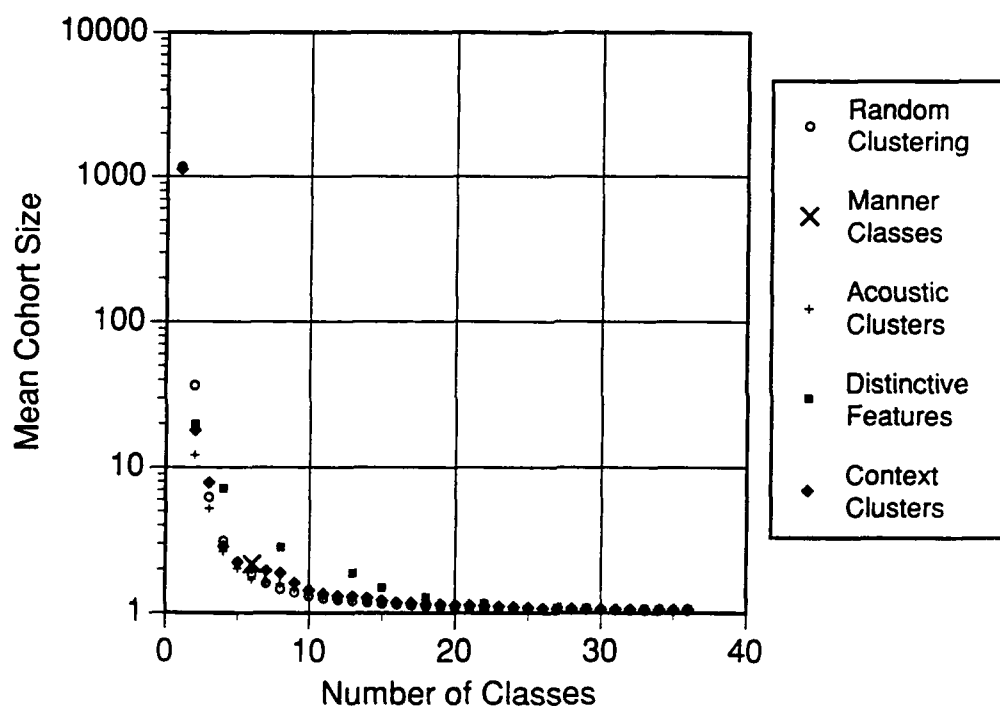


Figure B.4: Graph of phoneme class performance as measured by mean cohort size.

Bibliography

- [1] P. Ladefoged, *A Course in Phonetics*. Harcourt Brace Jovanovitch, Inc., second ed., 1982.
- [2] E. C. Sagey, *The Representation of Features and Relations in Non-linear Phonology*. PhD thesis, Massachusetts Institute of Technology, 1986.
- [3] K. N. Stevens, "Phonetic features and lexical access," in *Proceedings, The Second Symposium on Advanced Man-Machine Interface Through Spoken Language*, 1988.
- [4] D. W. Shipman and V. W. Zue, "Properties of large lexicons: Implications for advanced isolated word recognition systems," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1982.
- [5] D. P. Huttenlocher, "Acoustic-phonetic and lexical constraints in word recognition: Lexical access using partial information," Master's thesis, Massachusetts Institute of Technology, 1984.
- [6] L. Fissore, P. Laface, G. Micca, and R. Pieraccini, "Interaction between fast lexical access and word verification in large vocabulary continuous speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1988.
- [7] D. M. Carter, "An information-theoretic analysis of phonetic dictionary access," *Computer Speech and Language* 2, 1987.
- [8] G.-J. Vernooij, G. Bloothoof, and Y. van Holsteijn, "A simulation study on the usefulness of broad phonetic classification in automatic speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1989.
- [9] C. E. Shannon, "Prediction and entropy of printed english," *Bell System Technical Journal*, January 1951.
- [10] F. Jelinek, "Self-organized language modeling for speech recognition." tech. rep., IBM Continuous Speech Recognition Group, 1985.

- [11] K. Church, W. Gale, P. Hanks, and D. Hindle, "Parsing, word associations and typical predicate-argument relations," in *Proceedings, Speech and Natural Language Workshop*, Defense Advanced Research Projects Agency, 1989.
- [12] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley and Sons, Inc., 1968.
- [13] M. A. Randolph, *Syllable-based Constraints on Properties of English Sounds*. PhD thesis, Massachusetts Institute of Technology, 1989.
- [14] L. F. Lamel, *Formalizing Knowledge used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops*. PhD thesis, Massachusetts Institute of Technology, 1988.
- [15] H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English*. Brown University Press, 1967.
- [16] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
- [17] K. N. Stevens, "Course handout, 6.541 speech communication," 1989.
- [18] J. R. Glass, *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*. PhD thesis, Massachusetts Institute of Technology, 1988.
- [19] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proceedings, Speech Recognition Workshop*. Defense Advanced Research Projects Agency, 1986.
- [20] J. E. Shoup, "American english orthographic-phonemic dictionary," Scientific Report AFOSR-TR-73-1143, Speech Communications Research Laboratory, Inc., 1973.
- [21] Illumind Unabridged, 571 Belden St., Ste. A, Monterey, CA 93940, *MobyPronunciator*. macintosh version 1.01 ed., 1989.